facebook

Fun with Shallow Models!

Deep Learning, May 1st 2017, New York University

Laurens van der Maaten

Introduction

- Are all great learning algorithms deep?

Introduction

- Are all great learning algorithms deep? No!

Introduction

- Are all great learning algorithms deep? No!

- Today, we will look at three shallow, yet powerful algorithms:
  - Language modeling (word2vec): Predict a word given its context
  - Text classification (fastText): Predict a label for a document
  - Visualization (t-SNE): Make a map of high-dimensional data

- We will also see that you should always run good baselines

Language models

Language models

- Language models aim to predict a word given its surrounding words

Language models

- Language models aim to predict a word given its surrounding words
- In other words, they aim to build a distribution $p(\mathcal{W}) = p(w_1, w_2, \ldots, w_{|\mathcal{W}|})$

Language models

- Language models aim to predict a word given its surrounding words
- In other words, they aim to build a distribution $p(\mathcal{W}) = p(w_1, w_2, \ldots, w_{|\mathcal{W}|})$

- Traditional language models are based on n-grams

Language models

- Language models aim to predict a word given its surrounding words
- In other words, they aim to build a distribution $p(\mathcal{W}) = p(w_1, w_2, \ldots, w_{|\mathcal{W}|})$

- Traditional language models are based on n-grams:
  - The likelihood of a sentence: $p(\mathcal{W}) = \prod_{w_i \in \mathcal{W}} p(w_i | w_{i-1}, w_{i-2}, \ldots, w_{i-n})$
  - All of the probabilities are obtained by counting over a large corpus:

$$p(w_i | w_{i-1}, w_{i-2}) = \frac{C(w_i, w_{i-1}, w_{i-2})}{C(w_{i-1}, w_{i-2})}$$

Representing discrete objects

- Represent each word as a vector of zeros, except for one element
- Set the value in the vector corresponding to the index in the set to 1:

```
vocabulary = (Monday, Tuesday, is, a, today)
Monday   = [          ]
Tuesday  = [          ]
is       = [          ]
a        = [          ]
today    = [          ]
```

* Example reproduced with permission from Mikolov.

Representing discrete objects

- Represent each word as a vector of zeros, except for one element
- Set the value in the vector corresponding to the index in the set to 1:

```
vocabulary = (Monday, Tuesday, is, a, today)
Monday   = [1 0 0 0 0]
Tuesday  = [0 1 0 0 0]
is       = [0 0 1 0 0]
a        = [0 0 0 1 0]
today    = [0 0 0 0 1]
```

* Example reproduced with permission from Mikolov.

Representing discrete objects

- Represent each word as a vector of zeros, except for one element
- Set the value in the vector corresponding to the index in the set to 1:

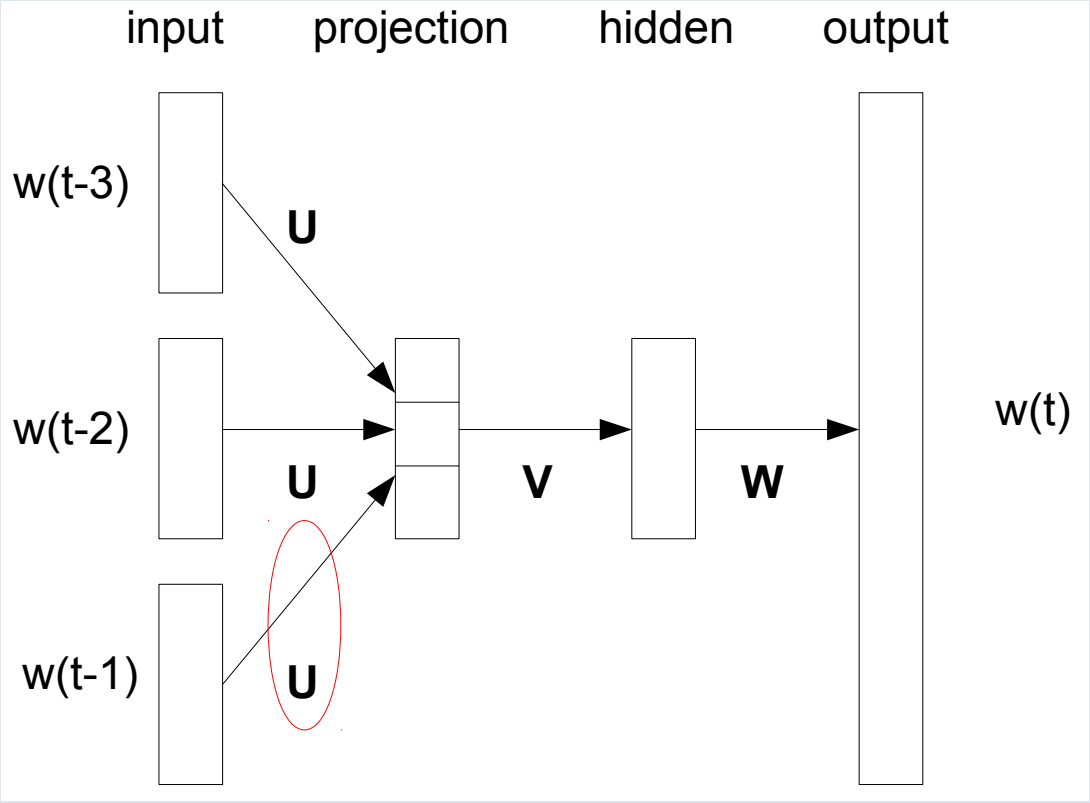
```
vocabulary = (Monday, Tuesday, is, a, today)
Monday   = [1 0 0 0 0]
Tuesday  = [0 1 0 0 0]
is       = [0 0 1 0 0]
a        = [0 0 0 1 0]
today    = [0 0 0 0 1]
```

- This is also known as a 1-of-K encoding (with K the vocabulary size)

* Example reproduced with permission from Mikolov.

A Feedforward Language Model

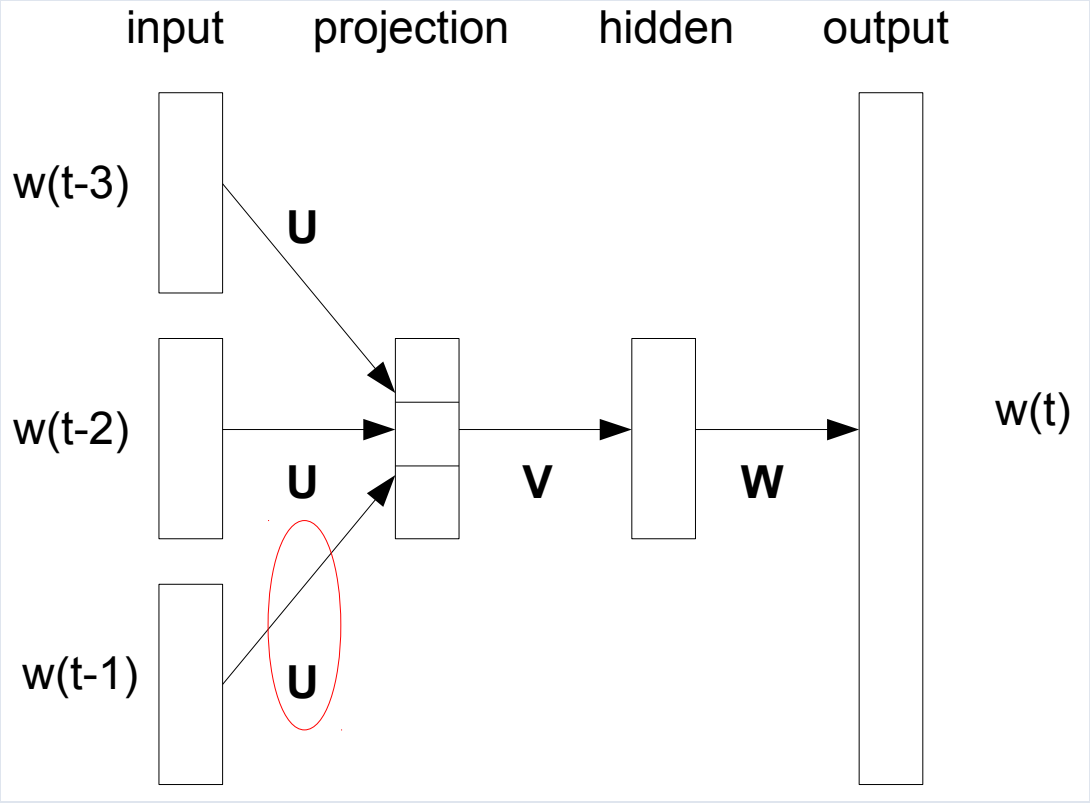
- We could use the following architecture for a word-prediction model:



* Figure reproduced with permission from Mikolov.

A Feedforward Language Model

- What does the matrix **U** actually model?



* Figure reproduced with permission from Mikolov.

Embedding

- Each column of **U** is an "embedding" of the corresponding word

Embedding

- Each column of **U** is an "embedding" of the corresponding word
- You can thus think of each word as being represented by a point that is "embedded" in a high-dimensional space

Embedding

- Each column of **U** is an "embedding" of the corresponding word
- You can thus think of each word as being represented by a point that is "embedded" in a high-dimensional space

- If the language model is trained well, then words that can be used interchangeably should have similar embeddings

Word2vec

- Word2vec is a very simple, very efficient language model:
  - It does not concatenate word embeddings, but it sums them
  - It does not use hidden layers
  - It shares weights between the input and output layer
  - It does not use a multi-class logistic loss (softmax) over predictions

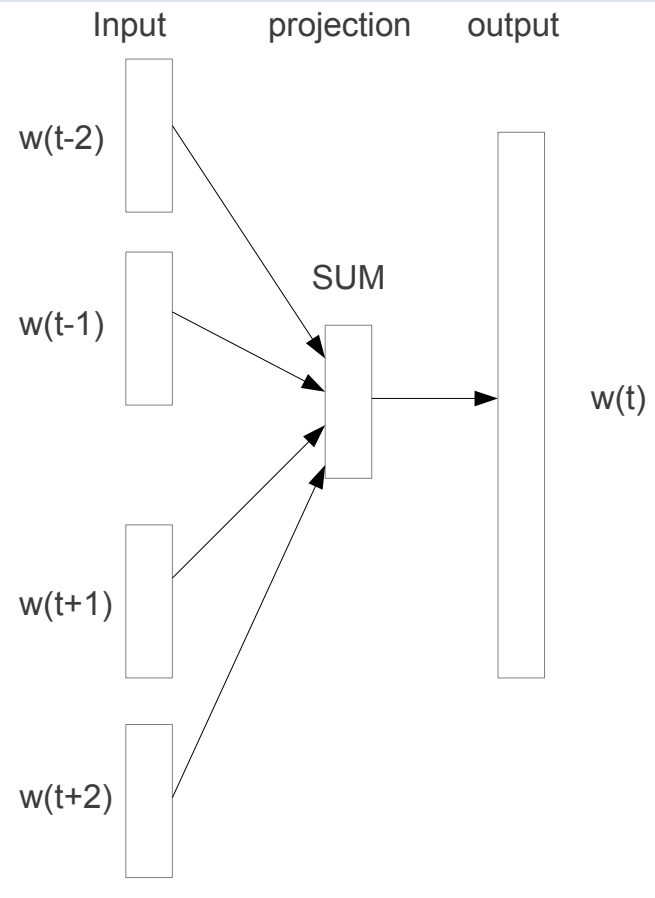
Word2vec

- Word2vec is a very simple, very efficient language model:
  - It does not concatenate word embeddings, but it sums them
  - It does not use hidden layers
  - It shares weights between the input and output layer
  - It does not use a multi-class logistic loss (softmax) over predictions

- Because it is so simple, it can be trained on billions of words
- This makes it scale much better than other language models

Word2vec: Architectures

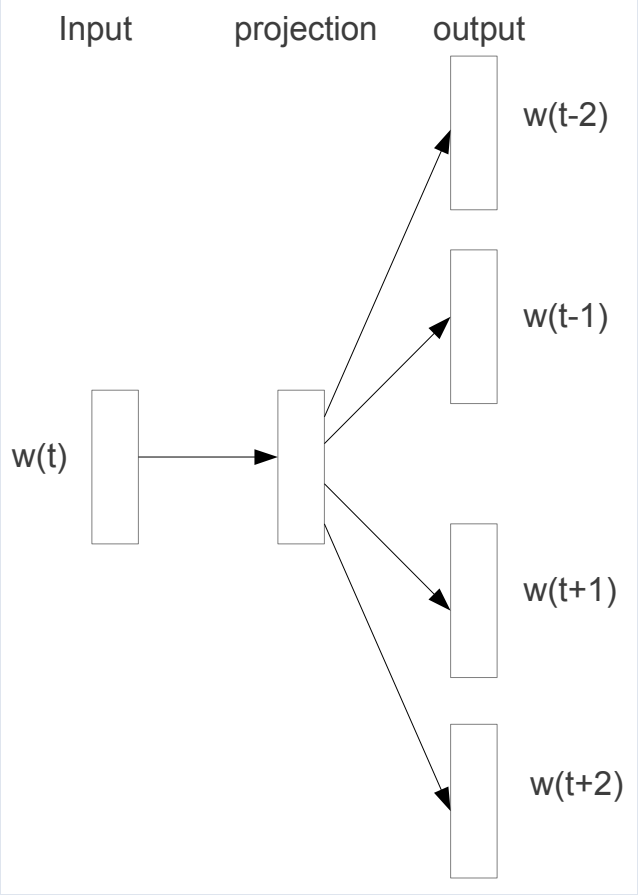
- "Continuous BoW" predicts current word given the surrounding words:



* Figure reproduced with permission from Mikolov.

Word2vec: Architectures

- "Skip-gram" predicts surrounding words given the current word:



* Figure reproduced with permission from Mikolov.

Loss function

- What loss are we optimizing in the word prediction?

Loss function

- What loss are we optimizing in the word prediction?

- A simple loss would be to minimize the multi-class logistic loss:

$$p(w_c|w_t) = \frac{\exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c})}{\sum_{w'} \exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w'})} \qquad \ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log p(w_c|w_t)$$

Loss function

- What loss are we optimizing in the word prediction?

- A simple loss would be to minimize the multi-class logistic loss:

$$p(w_c|w_t) = \frac{\exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c})}{\sum_{w'} \exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w'})} \qquad \ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log p(w_c|w_t)$$

softmax probability of predicting context word

Loss function

- What loss are we optimizing in the word prediction?

- A simple loss would be to minimize the multi-class logistic loss:

$$p(w_c|w_t) = \frac{\exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c})}{\sum_{w'} \exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w'})}$$

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log p(w_c|w_t)$$

(negative) log-probabilities summed over all words

Loss function

- What loss are we optimizing in the word prediction?

- A simple loss would be to minimize the multi-class logistic loss:

$$p(w_c|w_t) = \frac{\exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c})}{\sum_{w'} \exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w'})} \qquad \ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log p(w_c|w_t)$$

- Can you think of a problem when minimizing this loss?

Loss function

- What loss are we optimizing in the word prediction?

- A simple loss would be to minimize the multi-class logistic loss:

$$p(w_c|w_t) = \frac{\exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c})}{\sum_{w'} \exp(\mathbf{u}_{w_t}^\top \mathbf{u}_{w'})} \qquad \ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log p(w_c|w_t)$$

- Can you think of a problem when minimizing this loss?
  - The vocabulary may be very large! You will be waiting forever...

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \text{ with } v' \sim P(\mathcal{V})$$

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \text{ with } v' \sim P(\mathcal{V})$$

"positive" part

"negative" part

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \text{ with } v' \sim P(\mathcal{V})$$

encourage current word and context word to have a large inner product

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \text{ with } v' \sim P(\mathcal{V})$$

for all words in the context

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \quad \text{with} \quad \boxed{v' \sim P(\mathcal{V})}$$

pick a random word according to the unigram distribution

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \text{ with } v' \sim P(\mathcal{V})$$

encourage this random word to have a small inner product with the current word

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \text{ with } v' \sim P(\mathcal{V})$$

for K randomly picked words*

***NOTE: Sloppy notation alert. This is a different K than before!**

Word2vec: Loss function

- Word2vec minimizes a binary logistic loss on positive and negative samples:

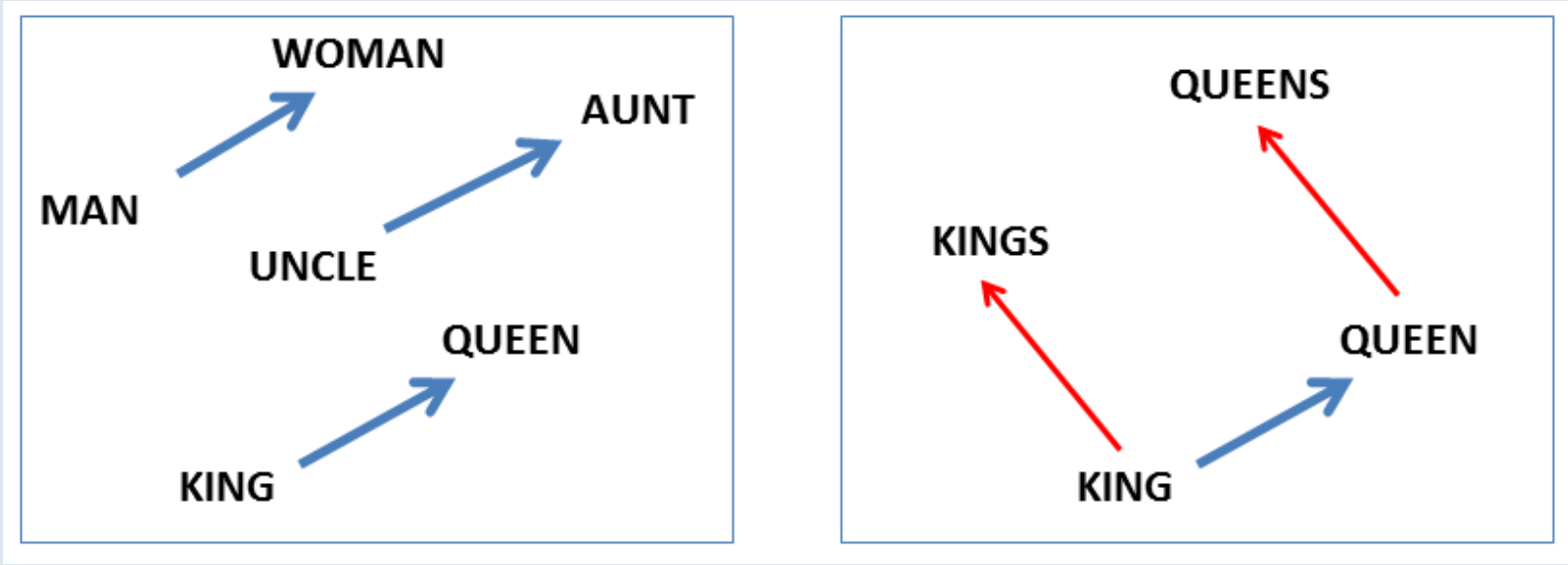
$$\ell(\mathbf{U}) = -\sum_{c \in \mathcal{C}} \log \sigma\left(\mathbf{u}_{w_t}^\top \mathbf{u}_{w_c}\right) - \sum_{j=1}^{K} \log \sigma\left(-\mathbf{u}_{w_t}^\top \mathbf{u}_{v'}\right) \text{ with } v' \sim P(\mathcal{V})$$

- Learning with SGD in this loss is very efficient:
  - Take a word and its context from a text corpus
  - Sample K words (typically, 5<K<20) from the unigram distribution
  - Compute the loss and the (sparse!) gradient update

- Multithreaded implementation allows for training speed of 5M words / sec

Linguistic regularities

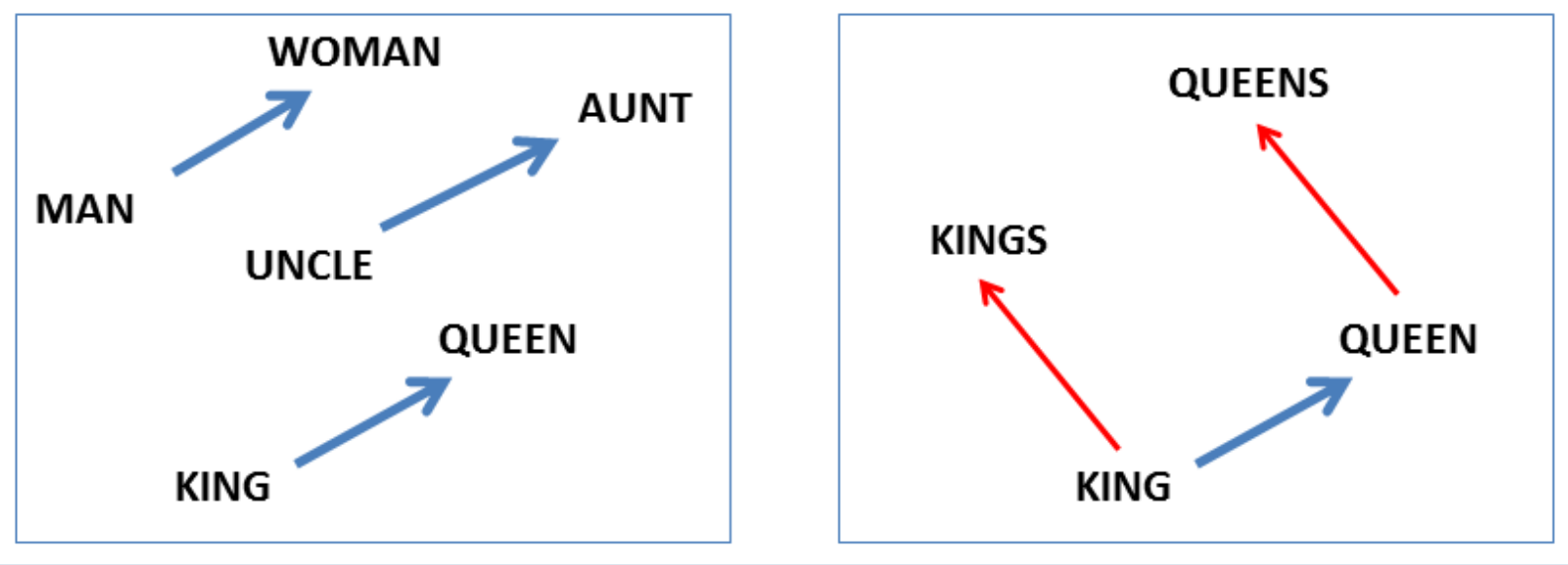
- Vector space implicitly encode regularities among words:



* Figure reproduced with permission from Mikolov.

Linguistic regularities

- Vector space implicitly encode regularities among words:



- We can exploit these regularities to do "linguistic arithmetic"

* Figure reproduced with permission from Mikolov.

Linguistic regularities

- Make dataset with analogies: "A is to B like C is to D"
- Answer the question "A is to B like C is to ?" by finding the word embedding that is closest to the embedding of B - A + C

Model	*Vector Dimensionality*	*Training Words*	*Training Time*	*Accuracy [%]*
Collobert NNLM	50	660M	2 months	11
Turian NNLM	200	37M	few weeks	2
Mnih NNLM	100	37M	7 days	9
Mikolov RNNLM	640	320M	weeks	25
Huang NNLM	50	990M	weeks	13
Our NNLM	100	6B	2.5 days	51
Skip-gram (hier.s.)	1000	6B	hours	66
CBOW (negative)	300	1.5B	**minutes**	**72**

* Table reproduced with permission from Mikolov.

Linguistic regularities

- Some examples of regularities:

Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

* Table reproduced with permission from Mikolov.

Linguistic regularities

- Compositionally by vector addition:

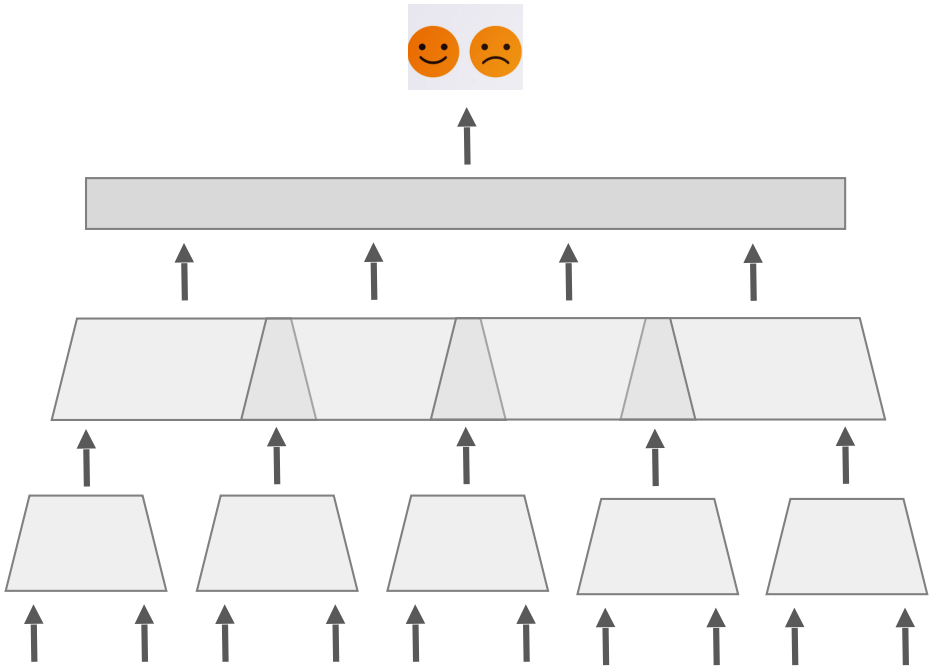
Expression	*Nearest tokens*
Czech + currency	koruna, Czech crown, Polish zloty, CTK
Vietnam + capital	Hanoi, Ho Chi Minh City, Viet Nam, Vietnamese
German + airlines	airline Lufthansa, carrier Lufthansa, flag carrier Lufthansa
Russian + river	Moscow, Volga River, upriver, Russia
French + actress	Juliette Binoche, Vanessa Paradis, Charlotte Gainsbourg

* Table reproduced with permission from Mikolov.

Text classification

Text classification

- Convolutional network
- State-of-the-art text classifiers (Zhang et al., 2015)
- What information do they capture?

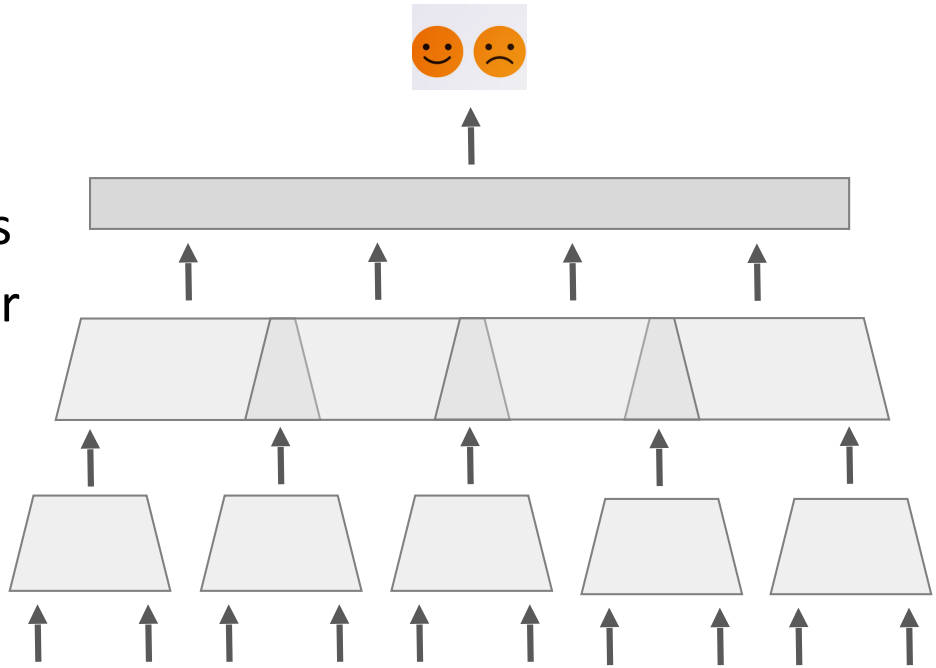


What a nice afternoon to enjoy my fav' band!

Slide credits: Armand Joulin

Text classification

- Convolutional networks:
  - Translation invariant
  - Information from groups of words
  - Separate embedding and classifier

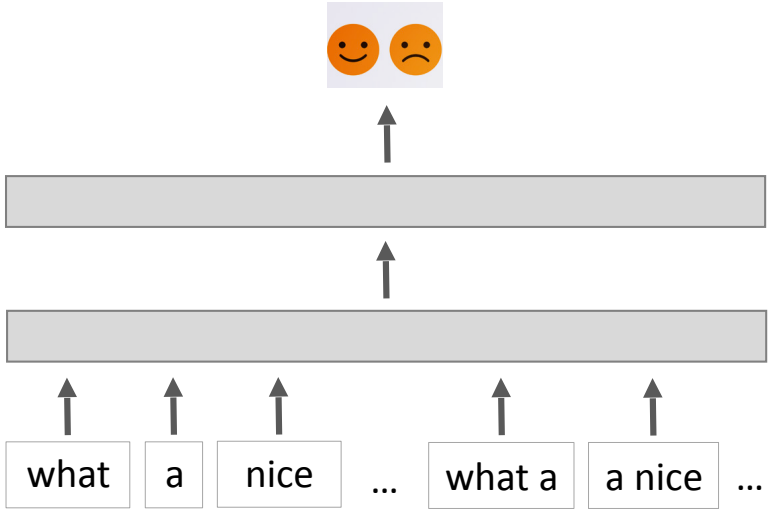


What a nice afternoon to enjoy my fav' band!

Slide credits: Armand Joulin

FastText

- n-grams features to capture interesting groups of words
- Low rank linear model to separate classifier/embedding
- Feature hashing to control model size



What a nice afternoon to enjoy my fav' band!

Slide credits: Armand Joulin

FastText

- Learning the following problem:

$$\min_{A \in \mathbb{R}^{m \times d},\ B \in \mathbb{R}^{d \times k}} -\frac{1}{N} \sum_{n=1}^{N} \ell(y_n, BAx_n)$$

- For efficiency, we use stochastic gradient descent:

$$A \leftarrow A - \lambda_t \nabla_A \ell(y_n, BAx_n)$$
$$B \leftarrow B - \lambda_t \nabla_B \ell(y_n, BAx_n)$$

- Asynchronous parallel updates with Hogwild (Recht, 2011)

Slide credits: Armand Joulin

FastText

- Choice of loss $\ell : (x, y) \mapsto \ell(x, y)$

- Standard loss is a log-softmax function:

$$\ell(x, y) = \log\left(\frac{\exp(y^T x)}{\sum_{k=1}^{K} \exp(x_k)}\right)$$

- The normalization term is linear in the number of classes

Slide credits: Armand Joulin

Comparison with convolutional networks

	Zhang et al. (2015)		Conneau et al. (2016)		fastText	
AG	87.2	3h	91.3	51m	92.5	1s
Amz. F.	59.5	5d	63.0	7h	60.2	9s
DBpedia	98.3	5h	98.7	1h	98.5	2s
Yah. A.	71.2	1d	73.4	2h	72.3	5s
Yelp F.	62.0	-	64.7	1h12	63.9	4s

Accuracy and train time

Same accuracy – **1k-10K times faster**

Slide credits: Armand Joulin

Always run some baselines!

Visual question answering

- Task: Answer a natural language question about an image

What color is the jacket?
-Red and blue.
-Yellow.
-Black.
-Orange.

How many cars are parked?
-Four.
-Three.
-Five.
-Six.

What event is this?
-A wedding.
-Graduation.
-A funeral.
-A picnic.

When is this scene taking place?
-Day time.
-Night time.
-Evening.
-Morning.

"State-of-the-art"

- Most current systems use a very similar approach:
  - Feed the **image** through a **convolutional network**
  - Feed the **question** to a **recurrent network** (LSTM), conditioning on the image features at initial state
  - Apply a **multi-class logistic regressor** on 7k frequent answers to terminal state of network
  - Train the convolutional network on Imagenet, and train the rest of the network **end-to-end**

"State-of-the-art"

- Most current systems use a very similar approach:
  - Feed the **image** through a **convolutional network**
  - Feed the **question** to a **recurrent network** (LSTM), conditioning on the image features at initial state
  - Apply a **multi-class logistic regressor** on 7k frequent answers to terminal state of network
  - Train the convolutional network on Imagenet, and train the rest of the network **end-to-end**

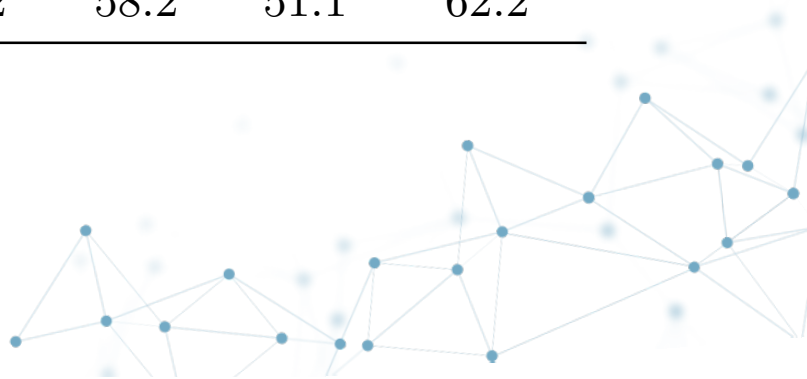
- Performance of some of these systems on Visual7W dataset:

Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2

A simple experiment

- Take a multiple-choice visual question

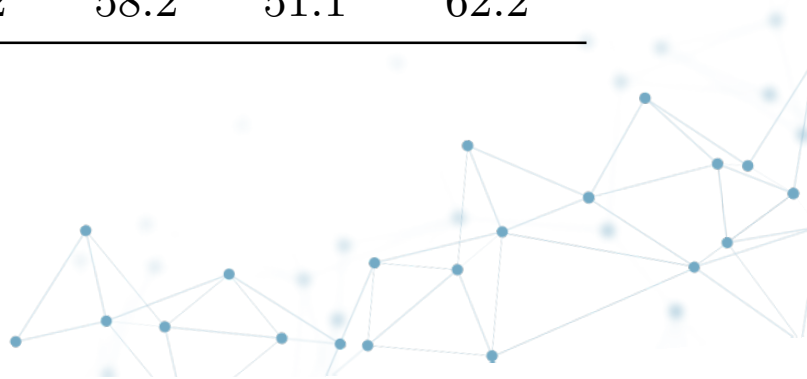
Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2



A simple experiment

- Take a multiple-choice visual question
- Throw away the image and the question

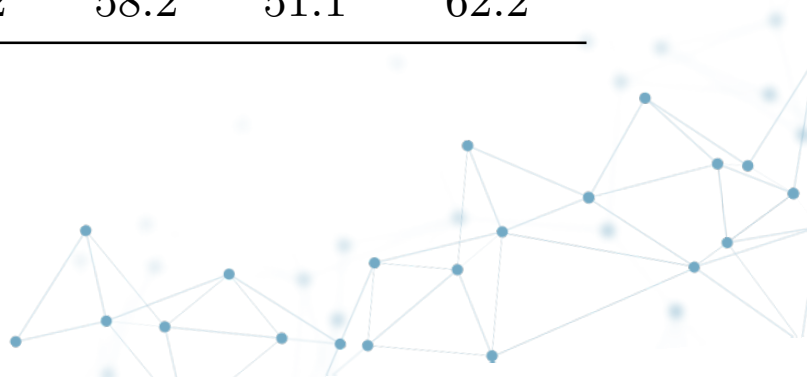
Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2



A simple experiment

- Take a multiple-choice visual question
- Throw away the image and the question
- Encode the answer using word2vec features (average over words)

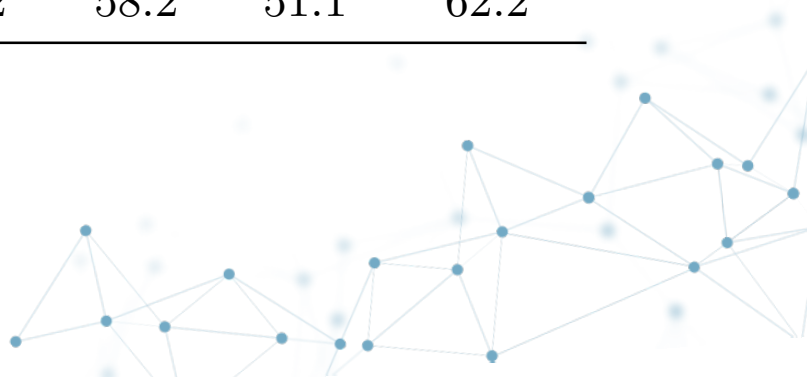
Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2



A simple experiment

- Take a multiple-choice visual question
- Throw away the image and the question
- Encode the answer using word2vec features (average over words)
- Train a binary classifier to predict whether or not the answer is correct

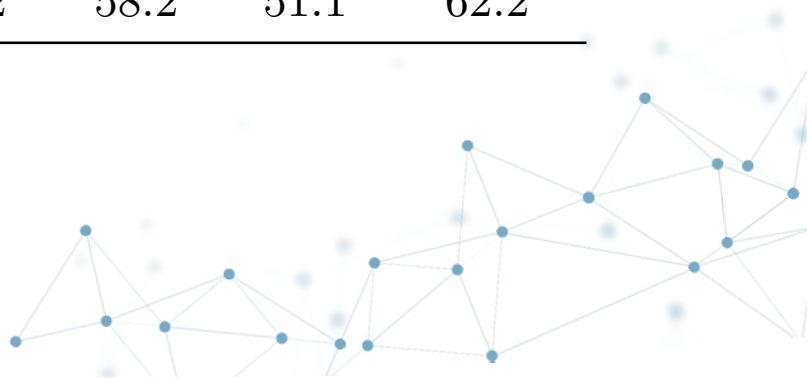
Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2



A simple experiment

- Take a multiple-choice visual question
- Throw away the image and the question
- Encode the answer using word2vec features (average over words)
- Train a binary classifier to predict whether or not the answer is correct
- At test time, predict the answer with the highest score

Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2



A simple experiment

- Take a multiple-choice visual question
- Throw away the image and the question
- Encode the answer using word2vec features (average over words)
- Train a binary classifier to predict whether or not the answer is correct
- At test time, predict the answer with the highest score

Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2
MLP (A)	47.3	58.2	74.3	63.6	57.1	49.6	52.9

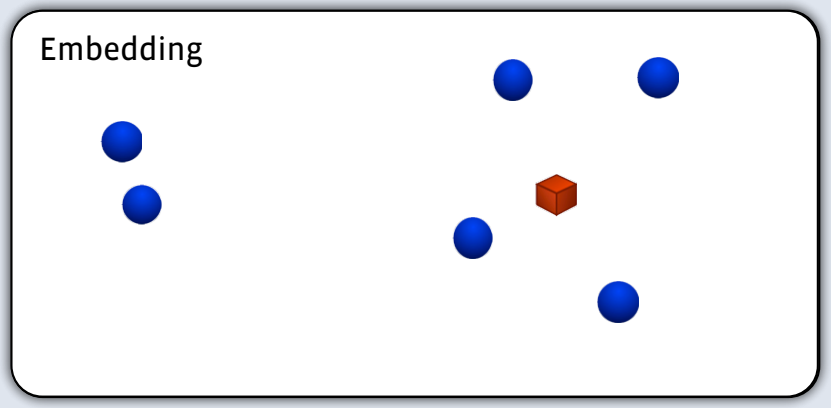
Clever Hans

Photo adopted from https://en.wikipedia.org/wiki/Clever_Hans

Data visualization

Introduction

- Presume we are given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- How do we learn a representation (embedding) for the vertices $v \in \mathcal{V}$?
  - We would like vertices that are connected to have similar embeddings
  - These embeddings we could then use in learning deep networks
  - If the embeddings are low-dimensional: use them to visualize the graph!



t-Stochastic Neighbor Embedding

- Convert the graph to a probability distribution over vertices:

$$p_{ij} = \frac{\exp(e_{ij})}{\sum_{k \neq l} \exp(e_{kl})}$$

- Strongly connected nodes will have large probabilities p_{ij}

t-Stochastic Neighbor Embedding

- Convert the graph to a probability distribution over vertices:

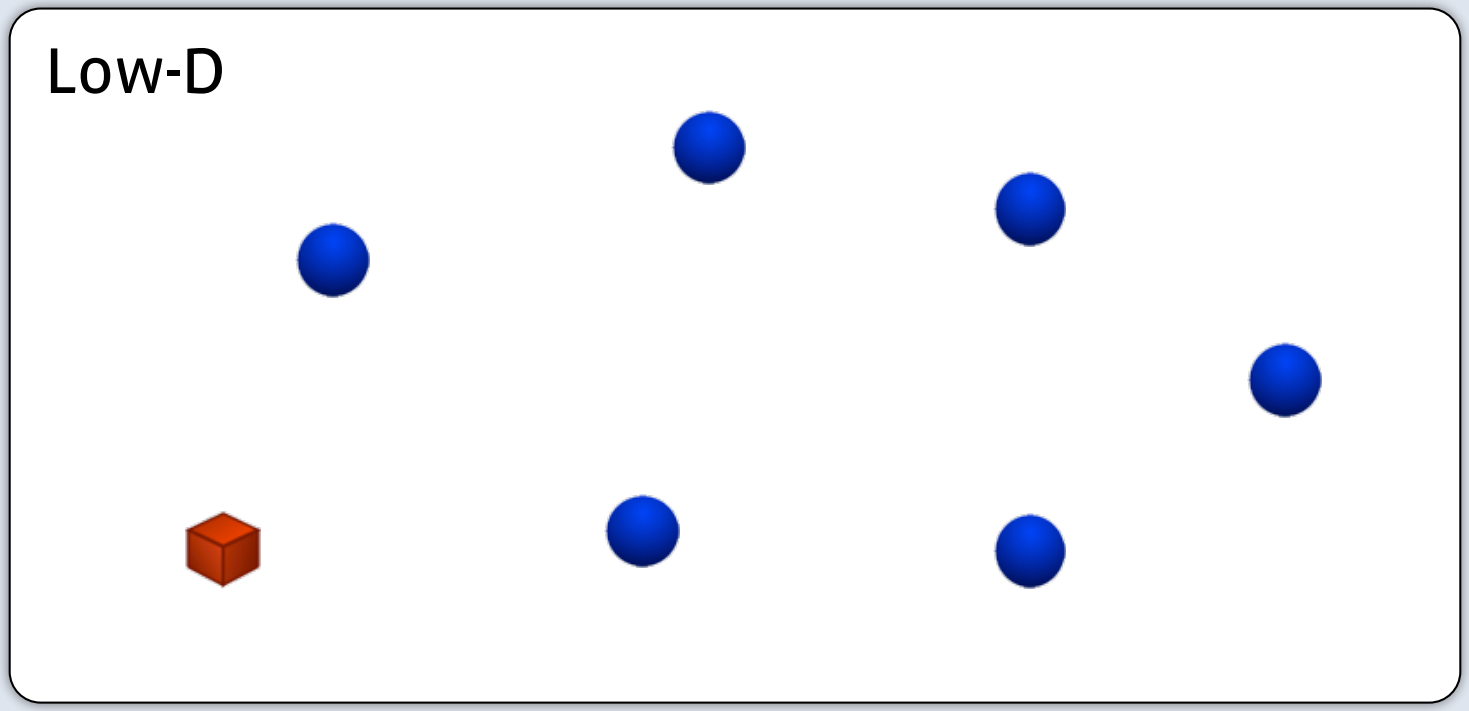
$$p_{ij} = \frac{\exp(e_{ij})}{\sum_{k \neq l} \exp(e_{kl})}$$

- Strongly connected nodes will have large probabilities p_{ij}

- When given high-dimensional data, we can compute similar probabilities:
  - For instance, set $e_{ij} = -\|\mathbf{x}_i - \mathbf{x}_j\|^2$ with $\mathbf{x}_i \in \mathbb{R}^D$
  - Note that this is essentially a (normalized) Gaussian kernel matrix

t-Stochastic Neighbor Embedding

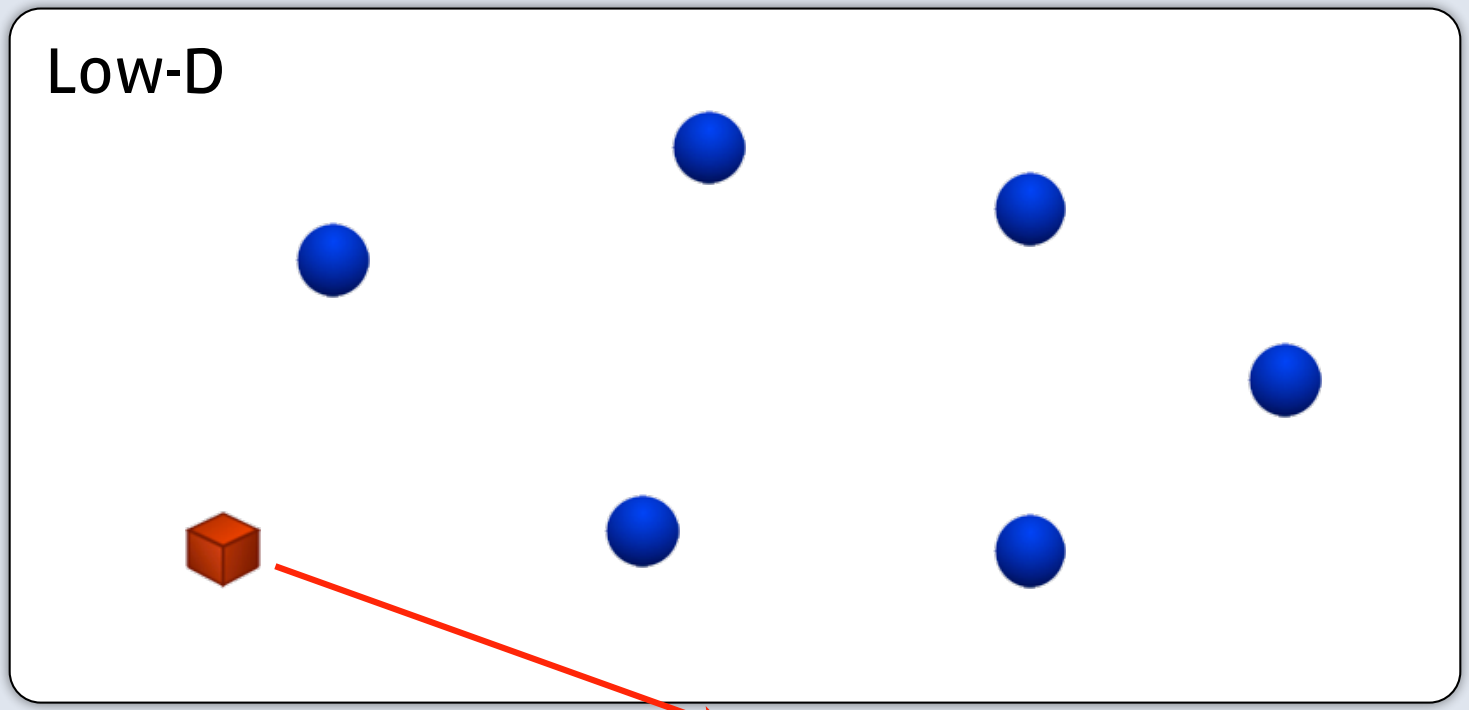
- Measure pairwise similarities between the embeddings:



$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

t-Stochastic Neighbor Embedding

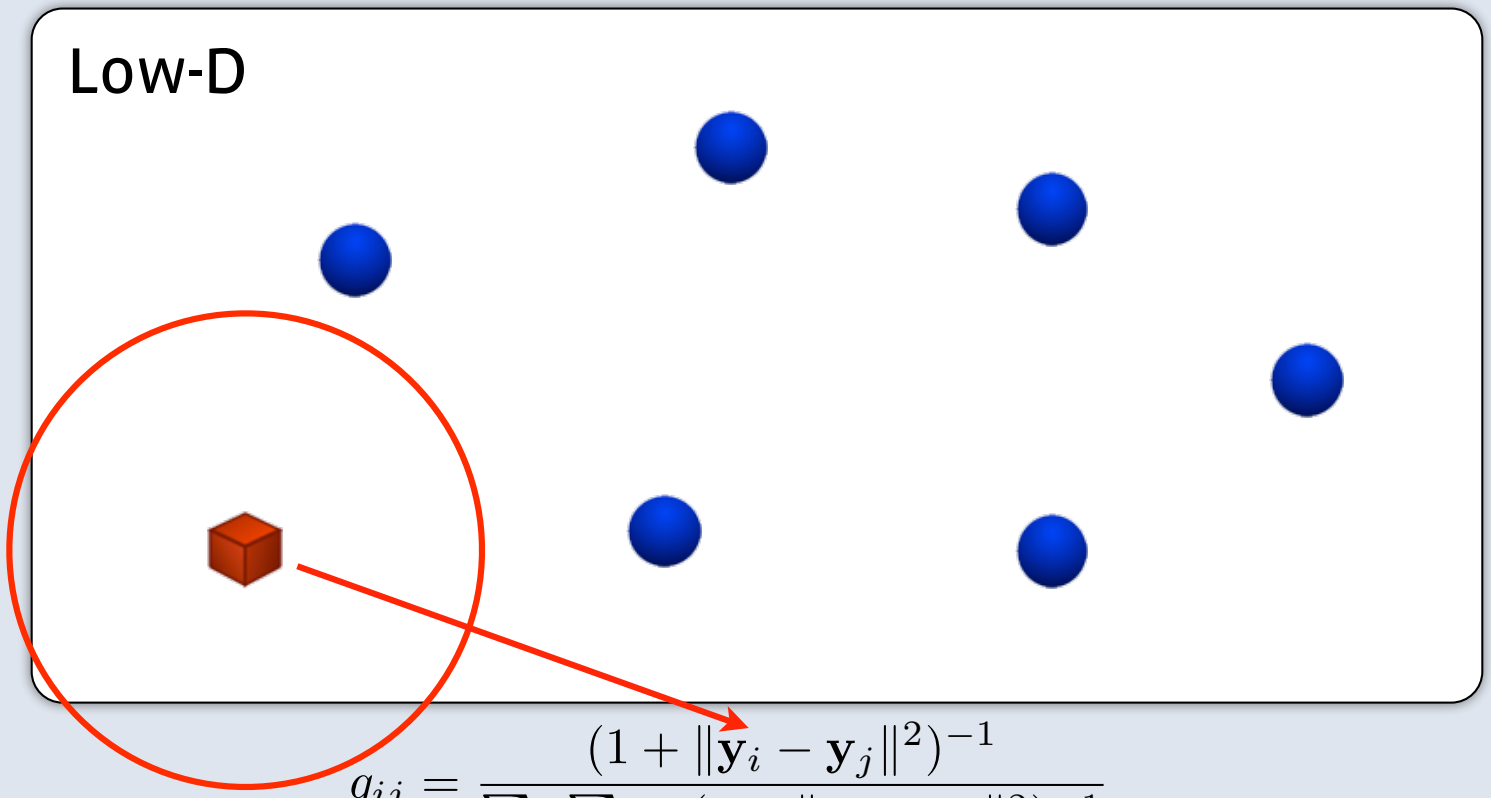
- Measure pairwise similarities between the embeddings:



$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

t-Stochastic Neighbor Embedding

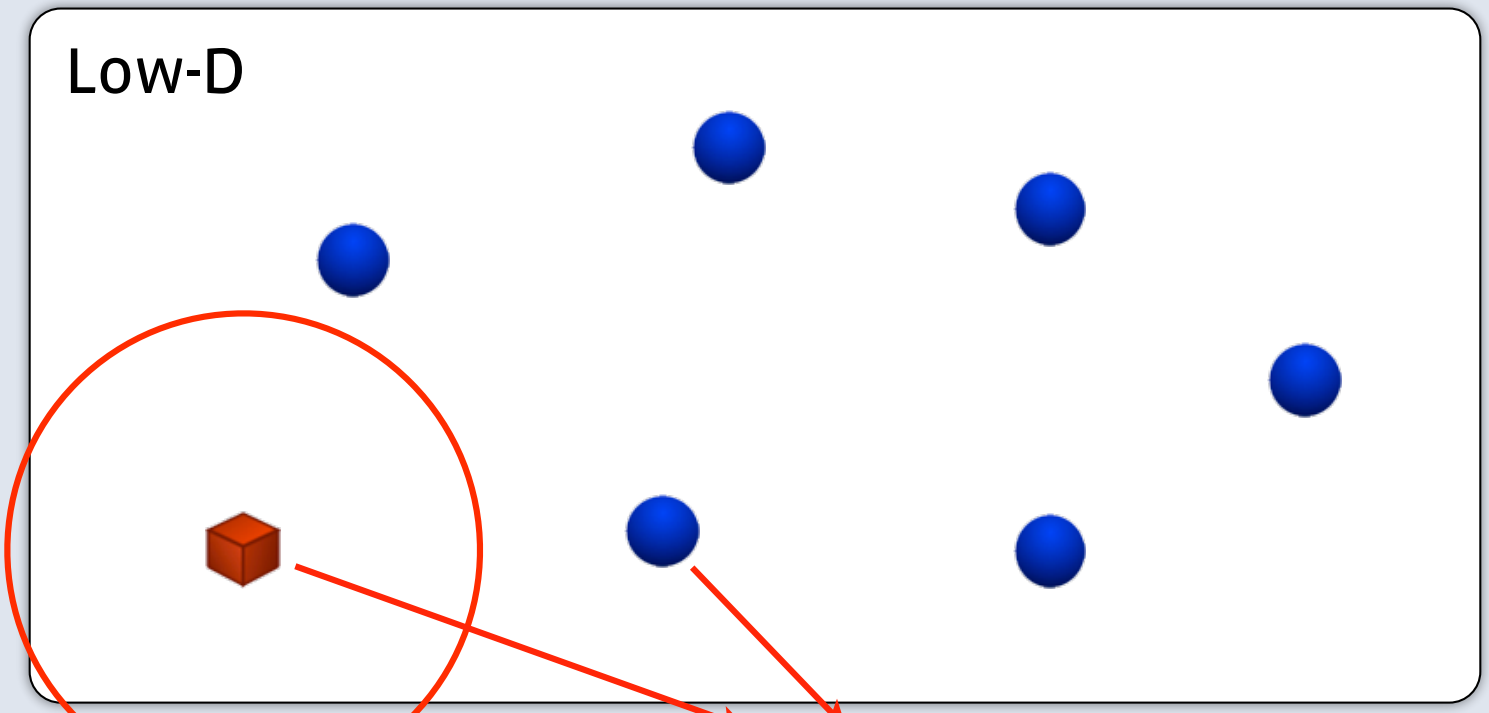
- Measure pairwise similarities between the embeddings:



$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

t-Stochastic Neighbor Embedding

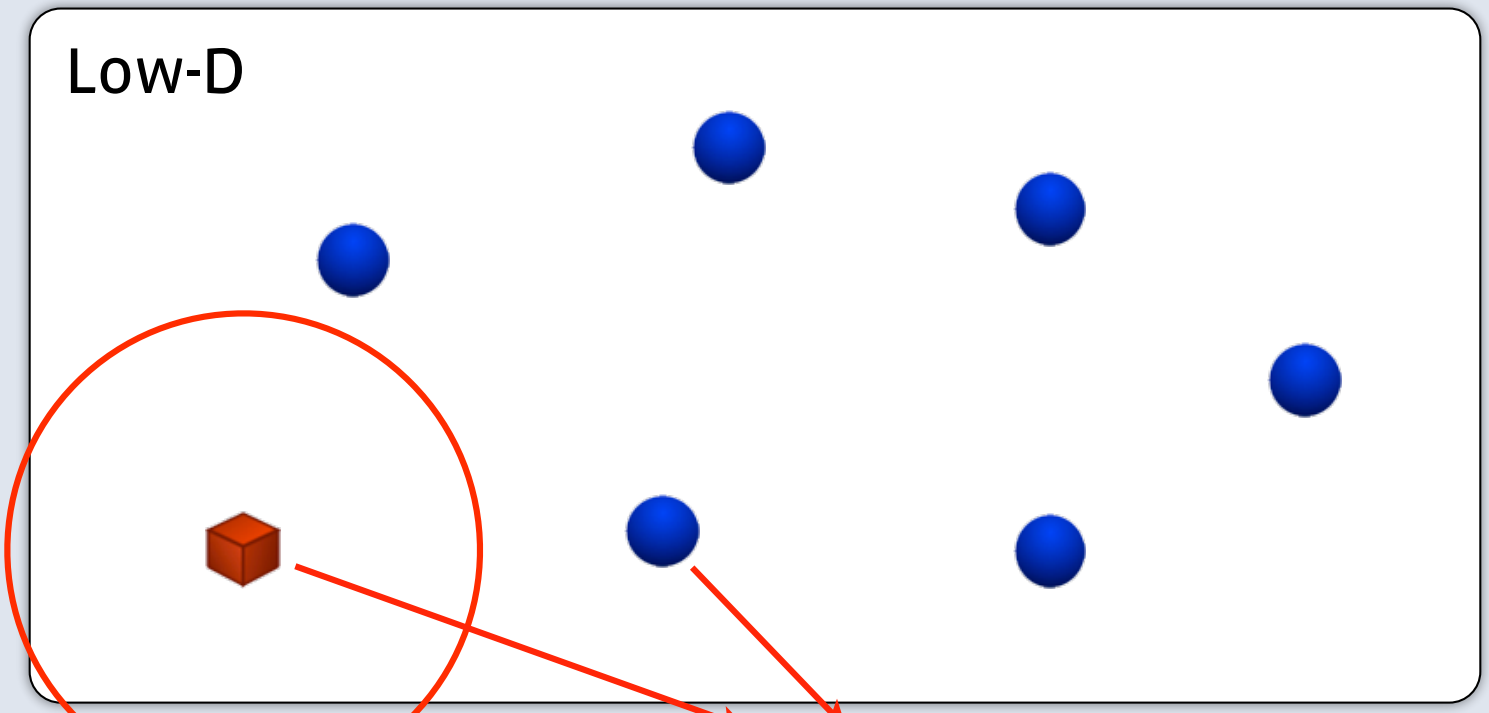
- Measure pairwise similarities between the embeddings:



$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

t-Stochastic Neighbor Embedding

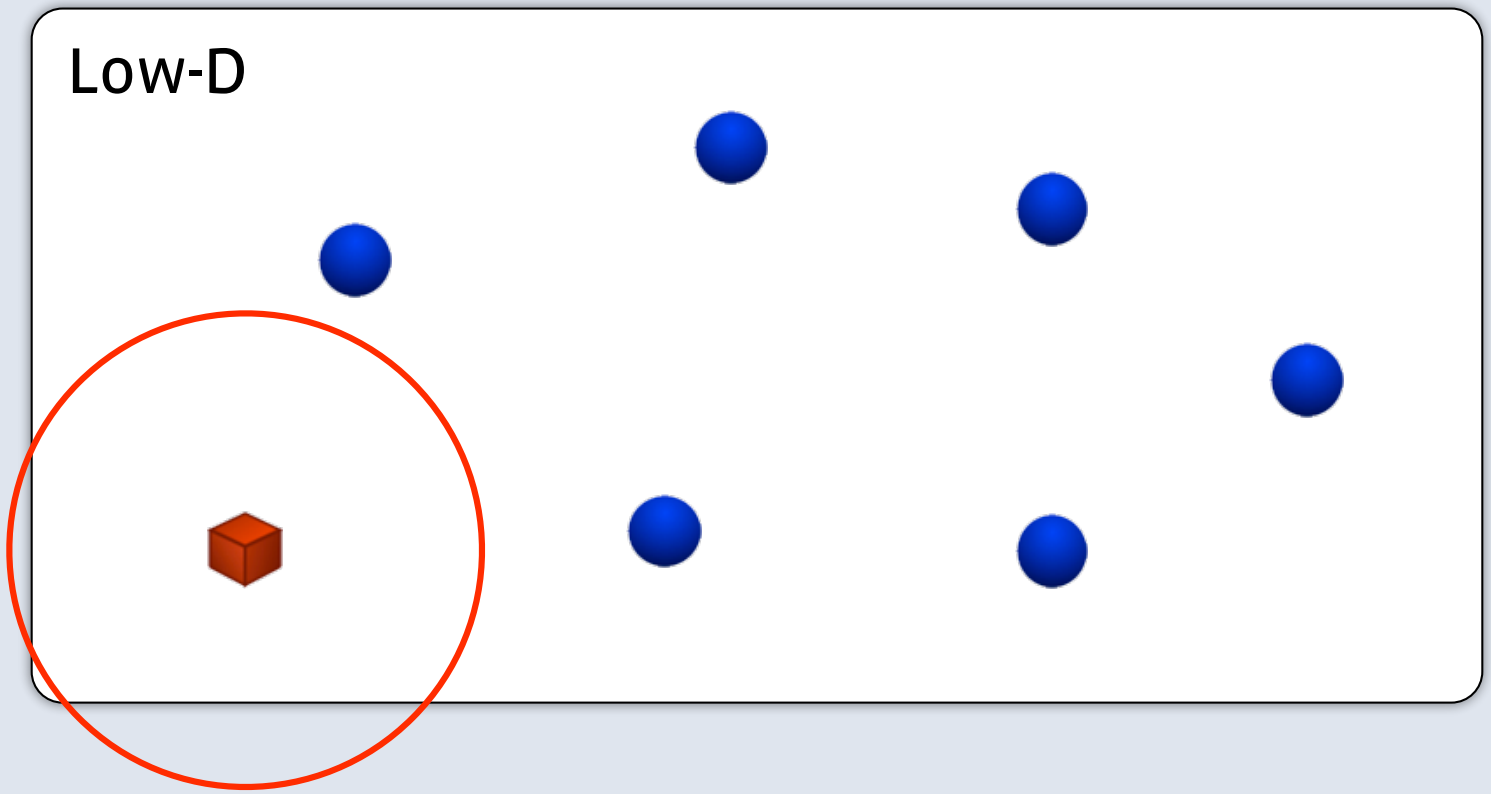
- Measure pairwise similarities between the embeddings:



$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

t-Stochastic Neighbor Embedding

- Move points around to minimize: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$



t-Stochastic Neighbor Embedding

- Kullback-Leibler divergence: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

t-Stochastic Neighbor Embedding

- Kullback-Leibler divergence: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$
  - Large p_{ij} modeled by small q_{ij}? Big penalty!
  - Small p_{ij} modeled by large q_{ij}? Not-so-big penalty.

t-Stochastic Neighbor Embedding

- Kullback-Leibler divergence: $KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$
  - Large p_{ij} modeled by small q_{ij}? Big penalty!
  - Small p_{ij} modeled by large q_{ij}? Not-so-big penalty.
- t-SNE makes sure connected vertices are close together in the embedding!

Visualization experiment

- Suppose we are given the MNIST dataset of handwritten digits

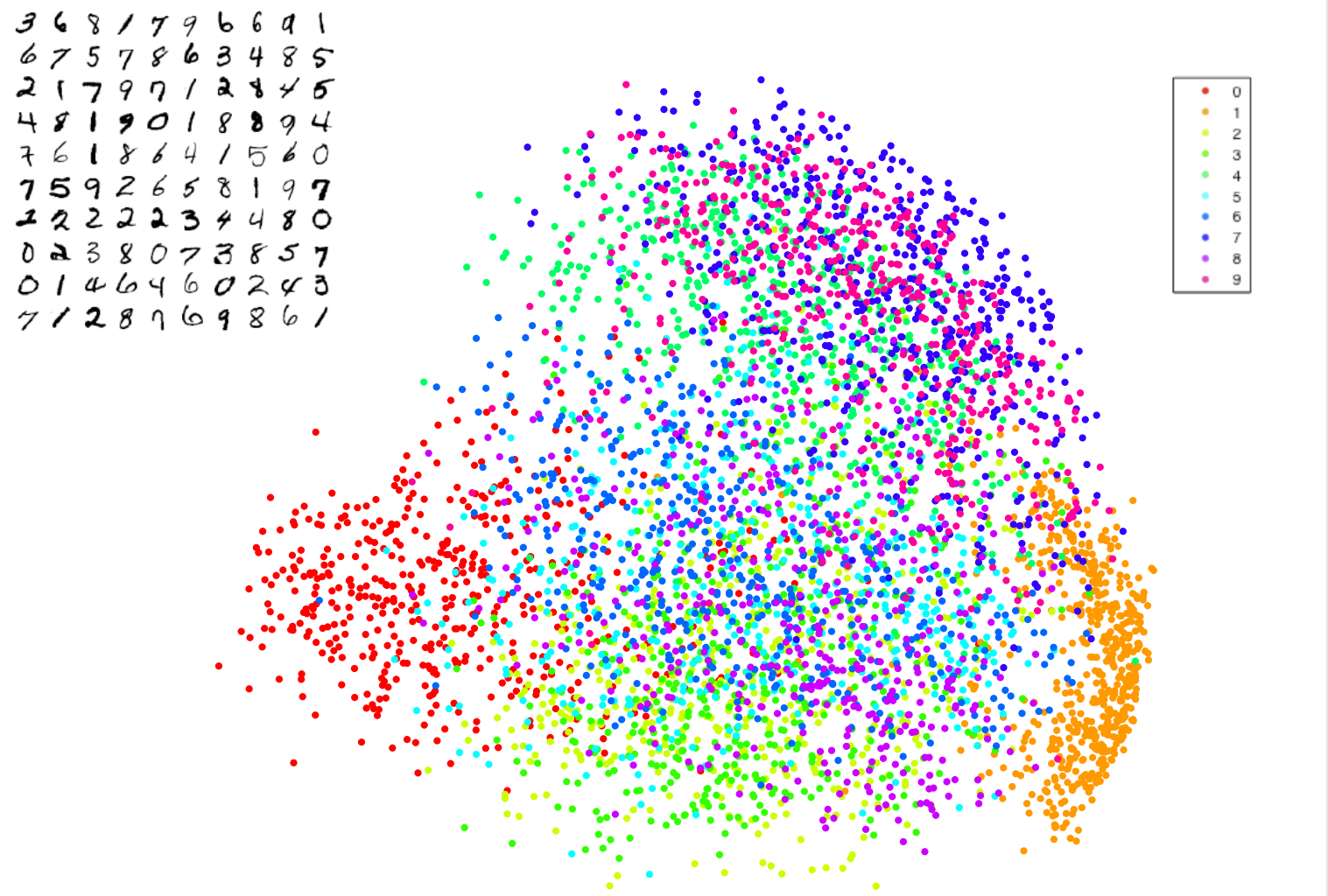
- Can we make a scatter plot that shows some of the structure in this data?

Visualization experiment

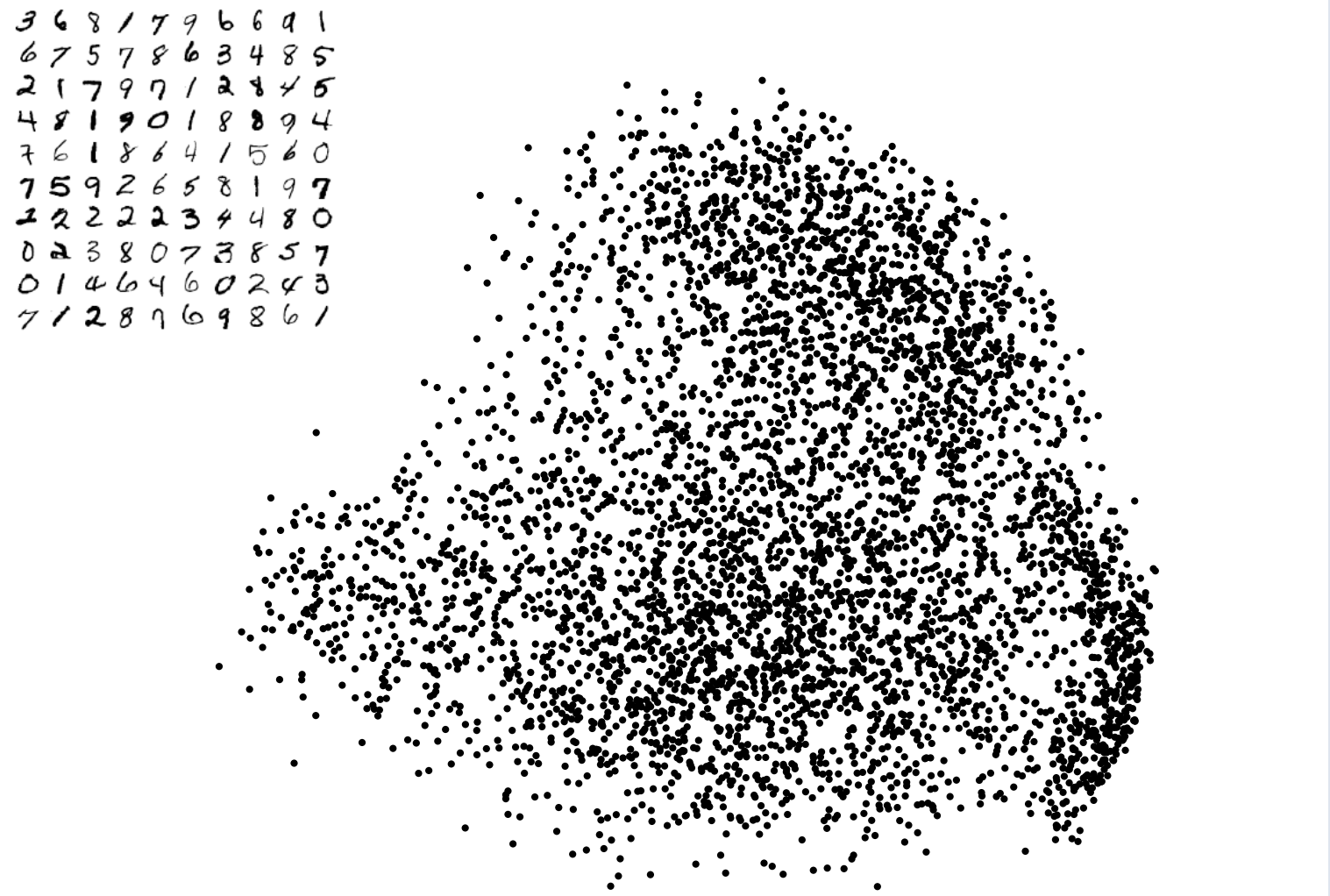
- Suppose we are given the MNIST dataset of handwritten digits

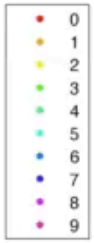
- Can we make a scatter plot that shows some of the structure in this data?
  - Approach 1:
    - Apply PCA on the digits and make a scatter plot in which the data is projected onto its first two principal components
  - Approach 2:
    - Construct a k-nearest neighbor graph (or simply compute a Gaussian kernel) and run t-SNE to learn a 2D embedding that you can show in a scatter plot

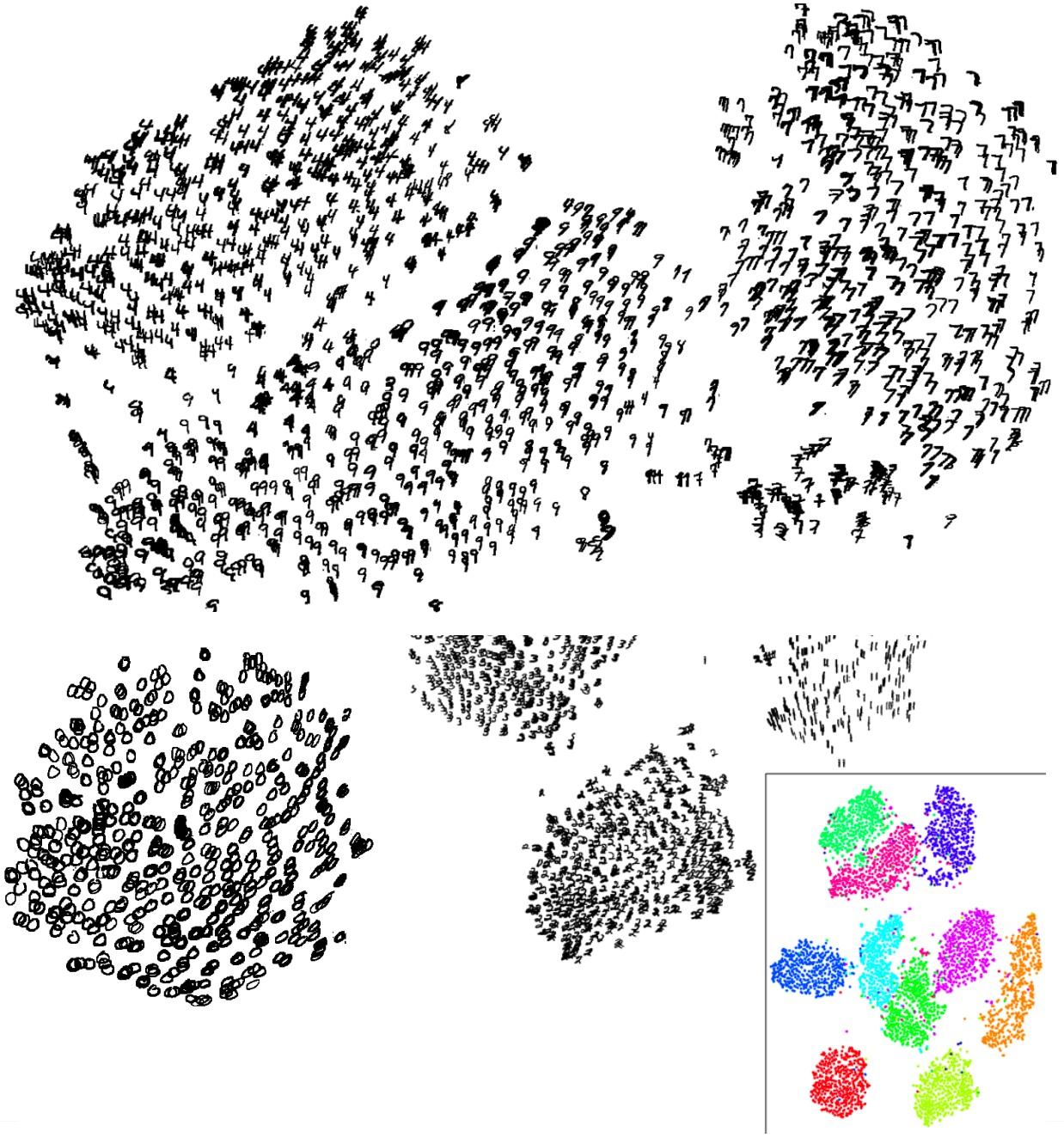
Principal Components Analysis

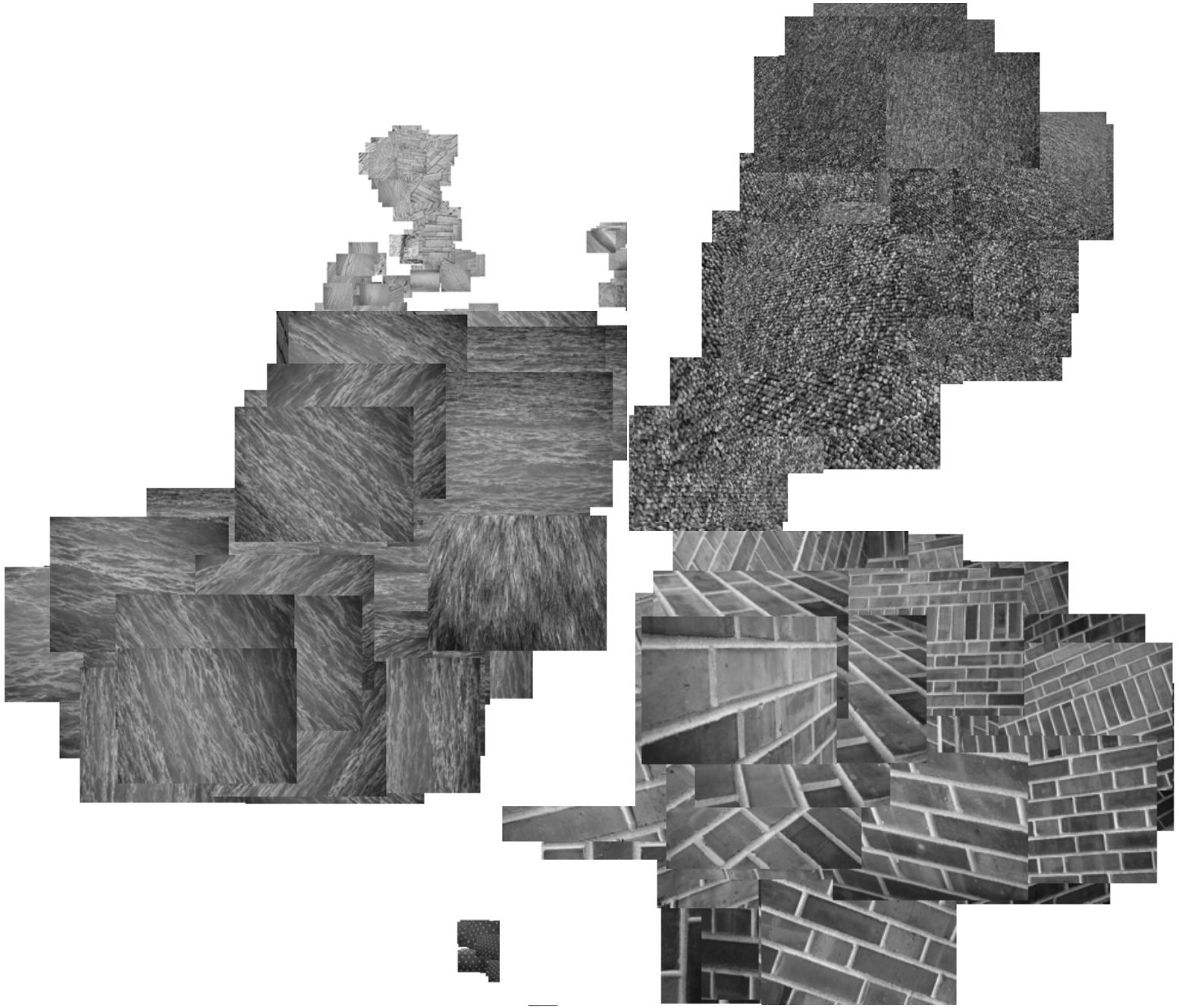


Principal Components Analysis



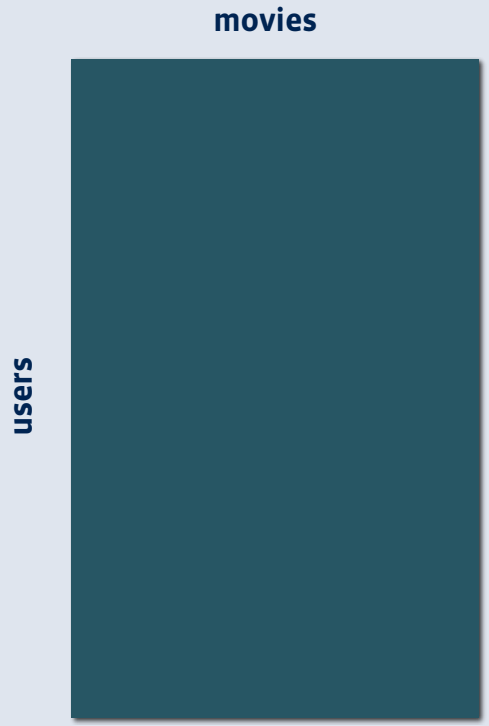

0
1
2
3
4
5
6
7
8
9





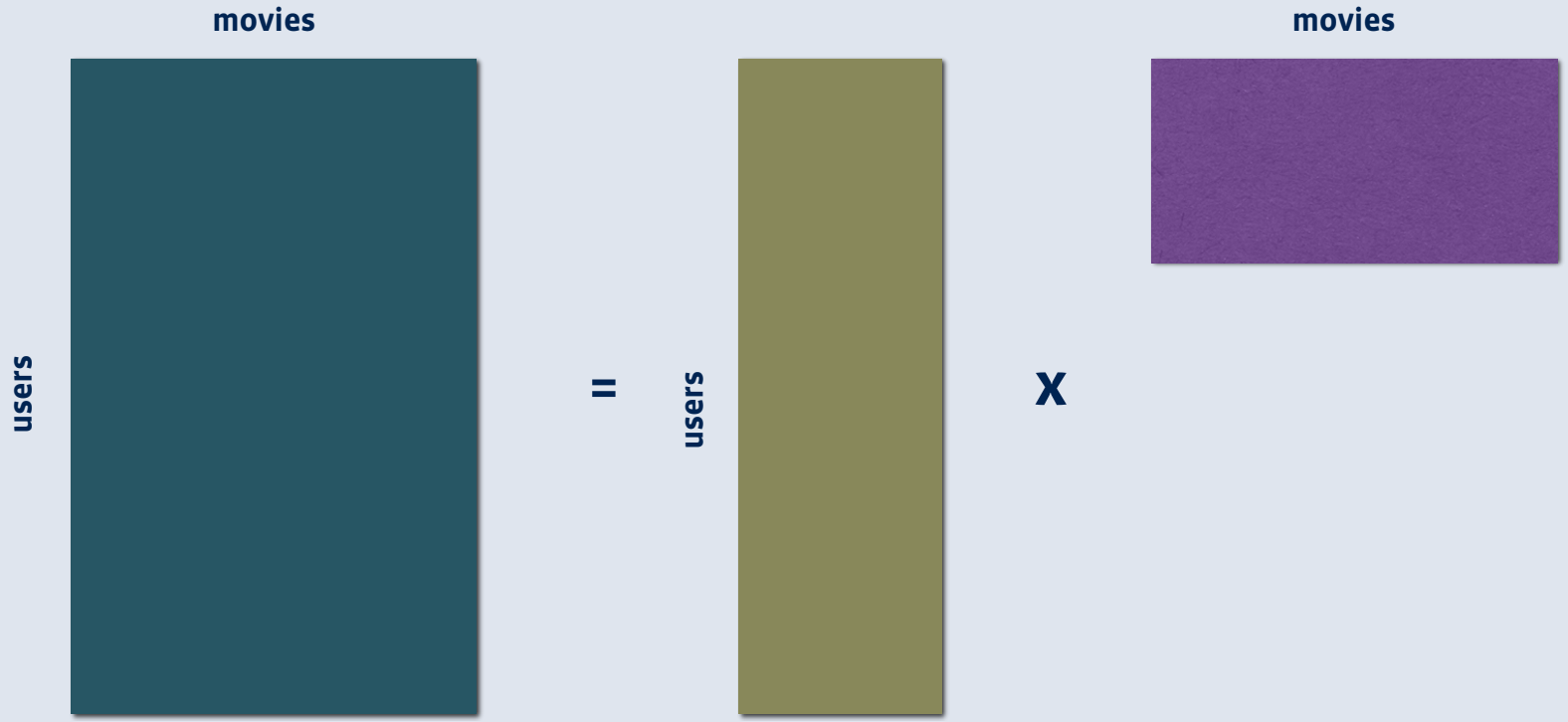
Visualizing movies

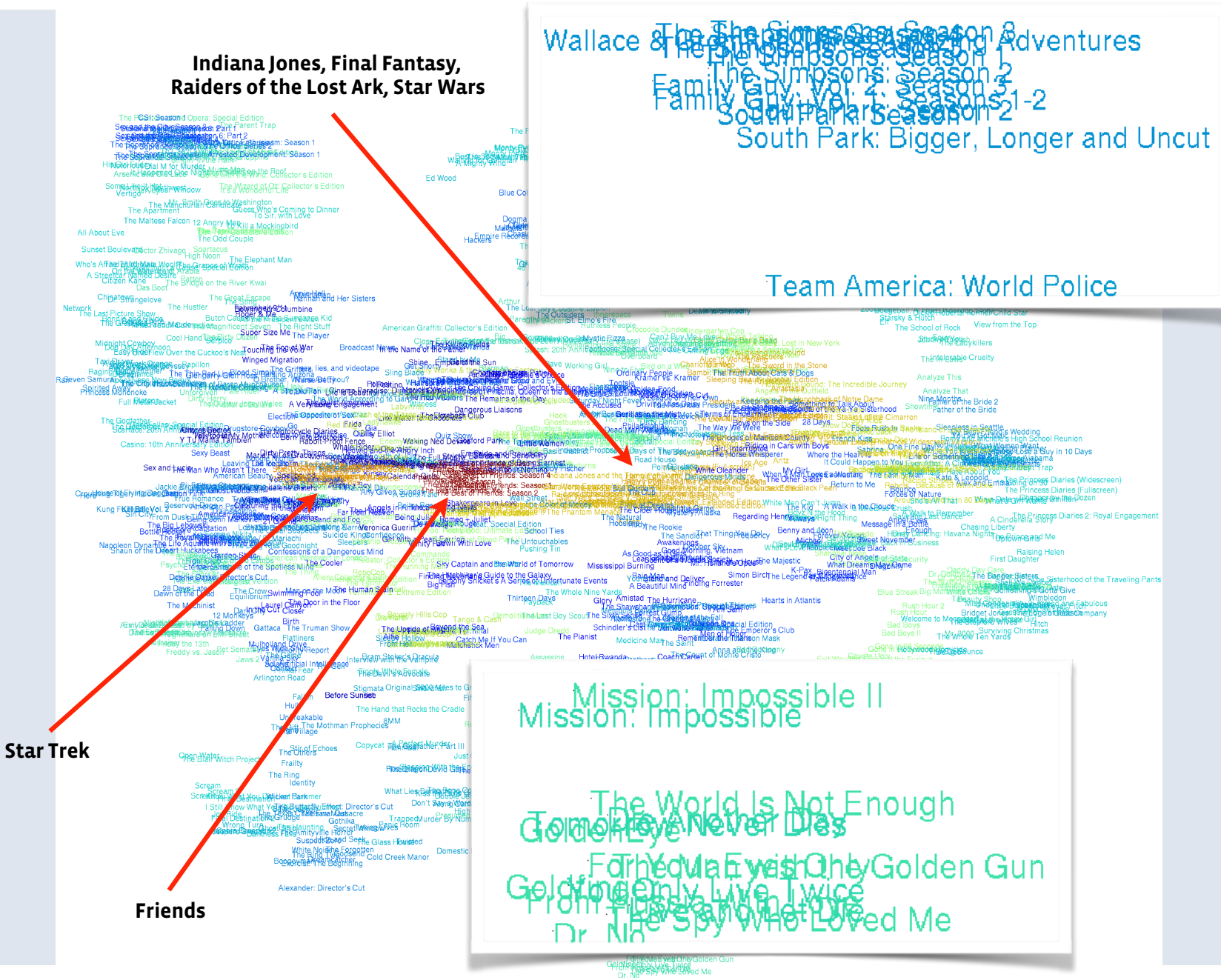
- Suppose you have a collection of user-movie ratings in a large *rating matrix*



Visualizing movies

- Suppose you have a collection of user-movie ratings in a large *rating matrix*
- *Decompose* the rating matrix to obtain *user features* and *movie features*:

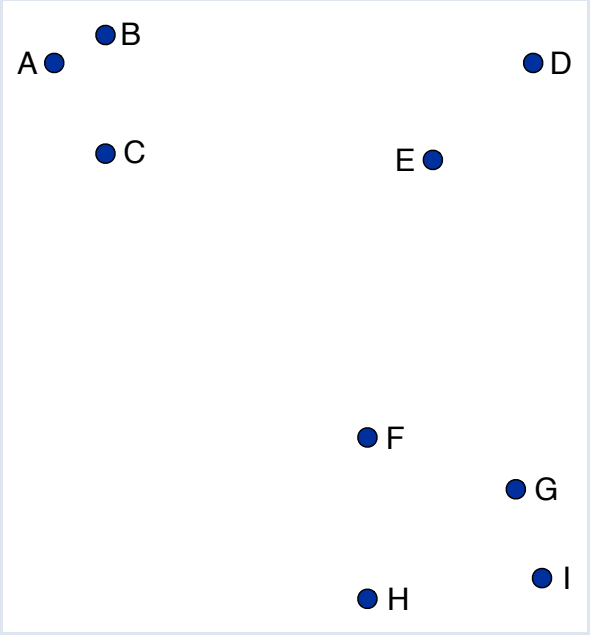



Indiana Jones, Final Fantasy, Raiders of the Lost Ark, Star Wars
Star Trek
Friends
Wallace & Gromit: The Best of Aardman Animation
The Simpsons: Season 8
The Simpsons: Season 1
The Simpsons: Season 2
Family Guy, Vol. 1: Seasons 1-2
South Park: Season 2
South Park: Bigger, Longer and Uncut
Team America: World Police
Mission: Impossible II
Mission: Impossible
The World Is Not Enough
Tomorrow Never Dies
GoldenEye
For Your Eyes Only
The Golden Gun
Goldfinger
Only Live Twice
From Russia With Love
The Spy Who Loved Me
Dr. No

Gradient interpretation

- We can interpret the t-SNE gradient as a simulation of an N-body system:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j)$$

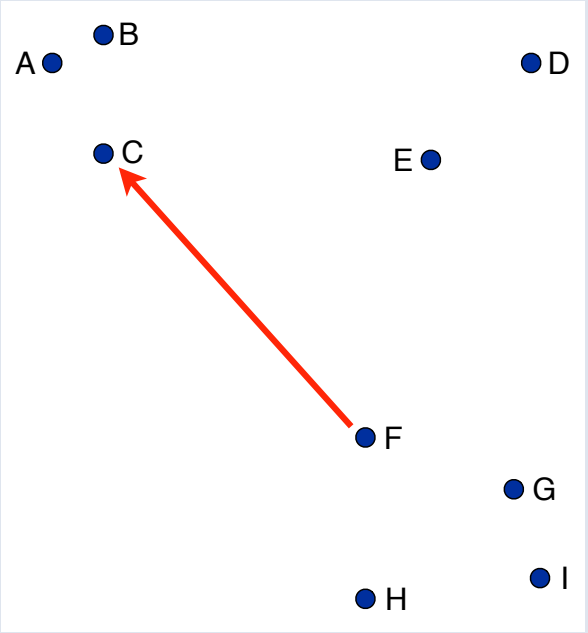


Gradient interpretation

- We can interpret the t-SNE gradient as a simulation of an N-body system:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j)$$

spring

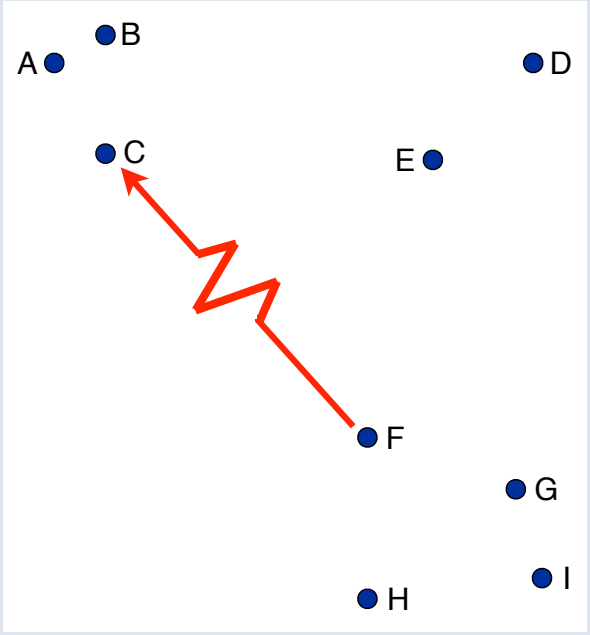


Gradient interpretation

- We can interpret the t-SNE gradient as a simulation of an N-body system:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j)$$

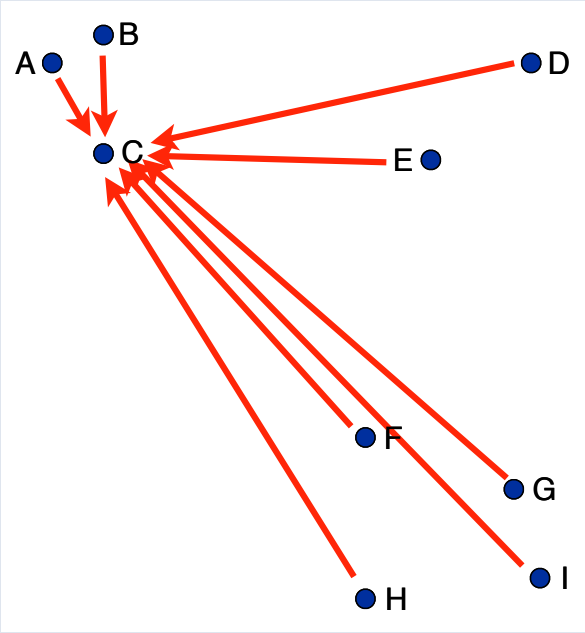
exertion / compression



Gradient interpretation

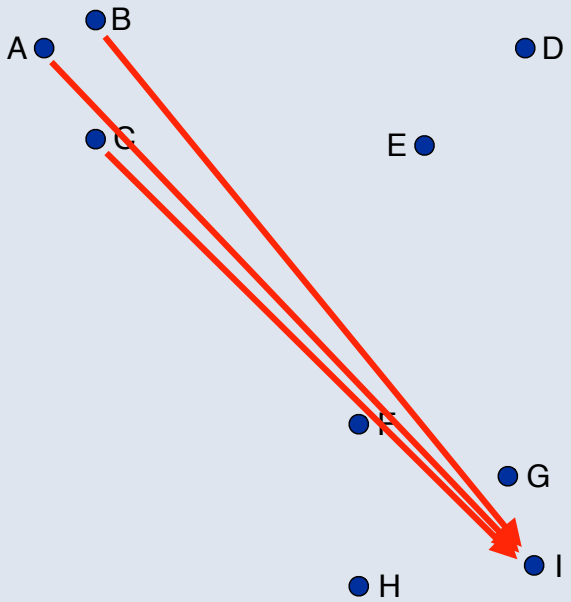
- We can interpret the t-SNE gradient as a simulation of an N-body system:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j)$$



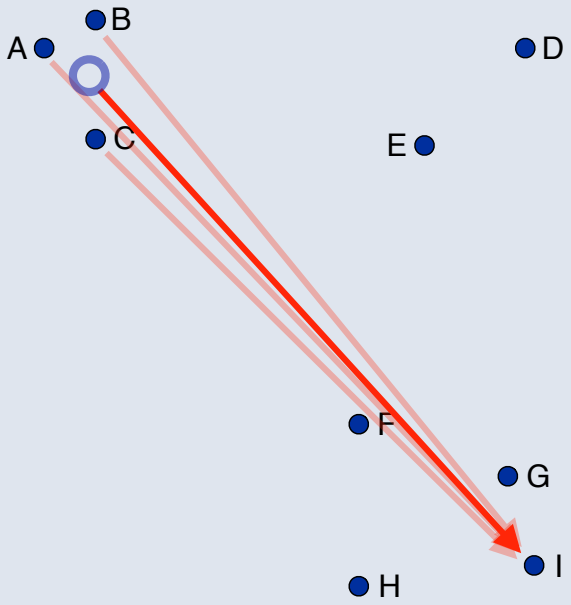
Barnes-Hut approximation

- Many of the pairwise interactions between points are very similar:



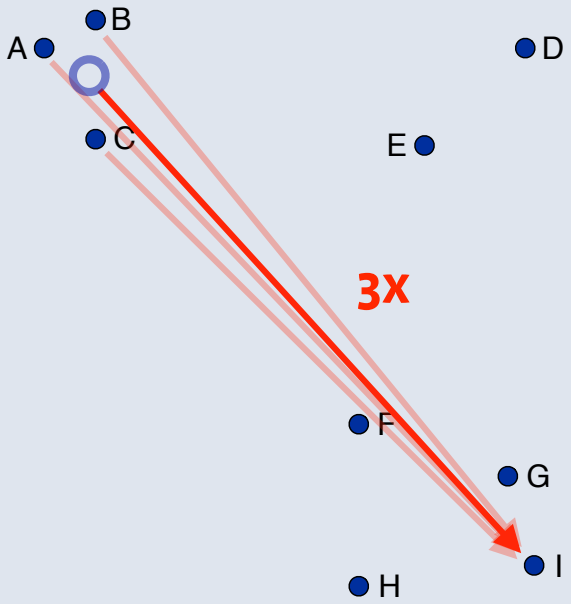
Barnes-Hut approximation

- Approximate such similar interactions by a single interaction:

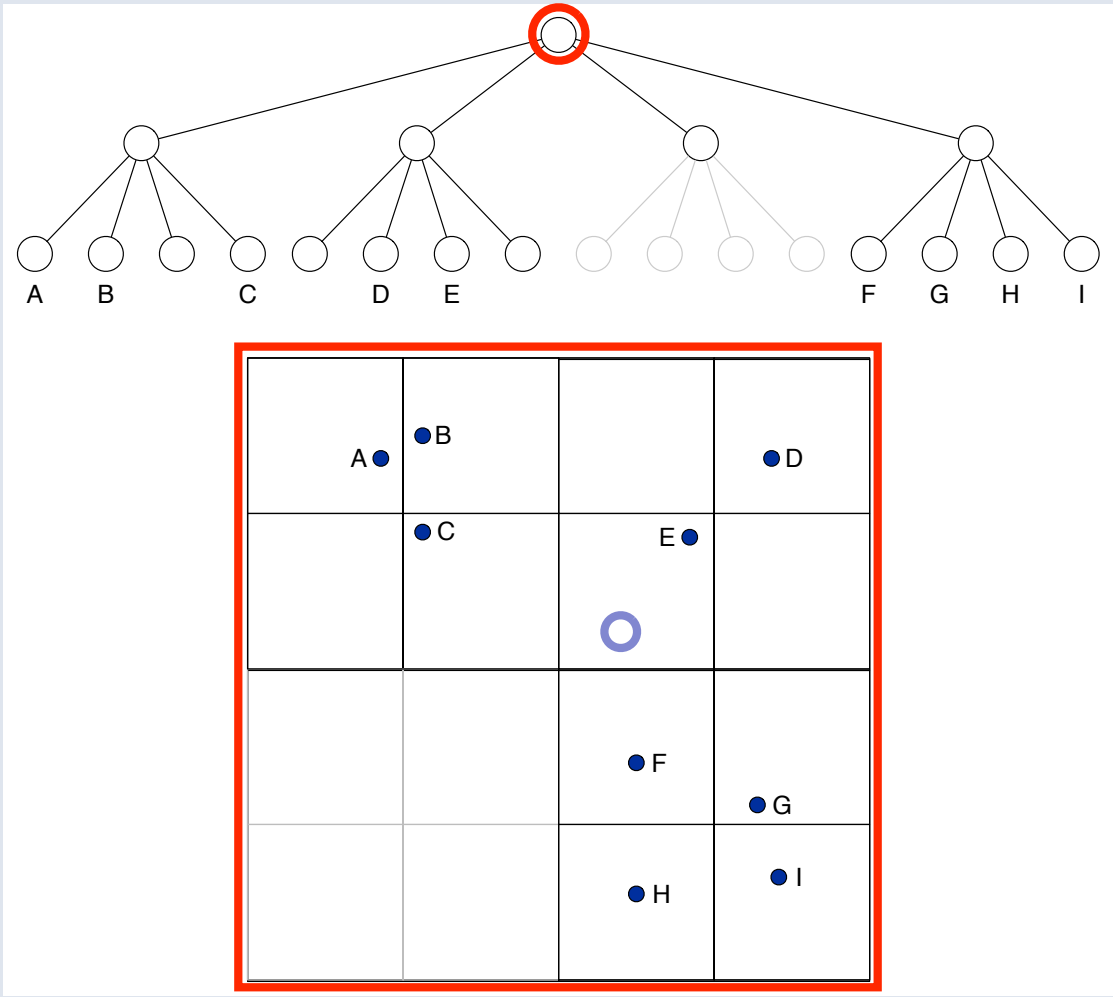


Barnes-Hut approximation

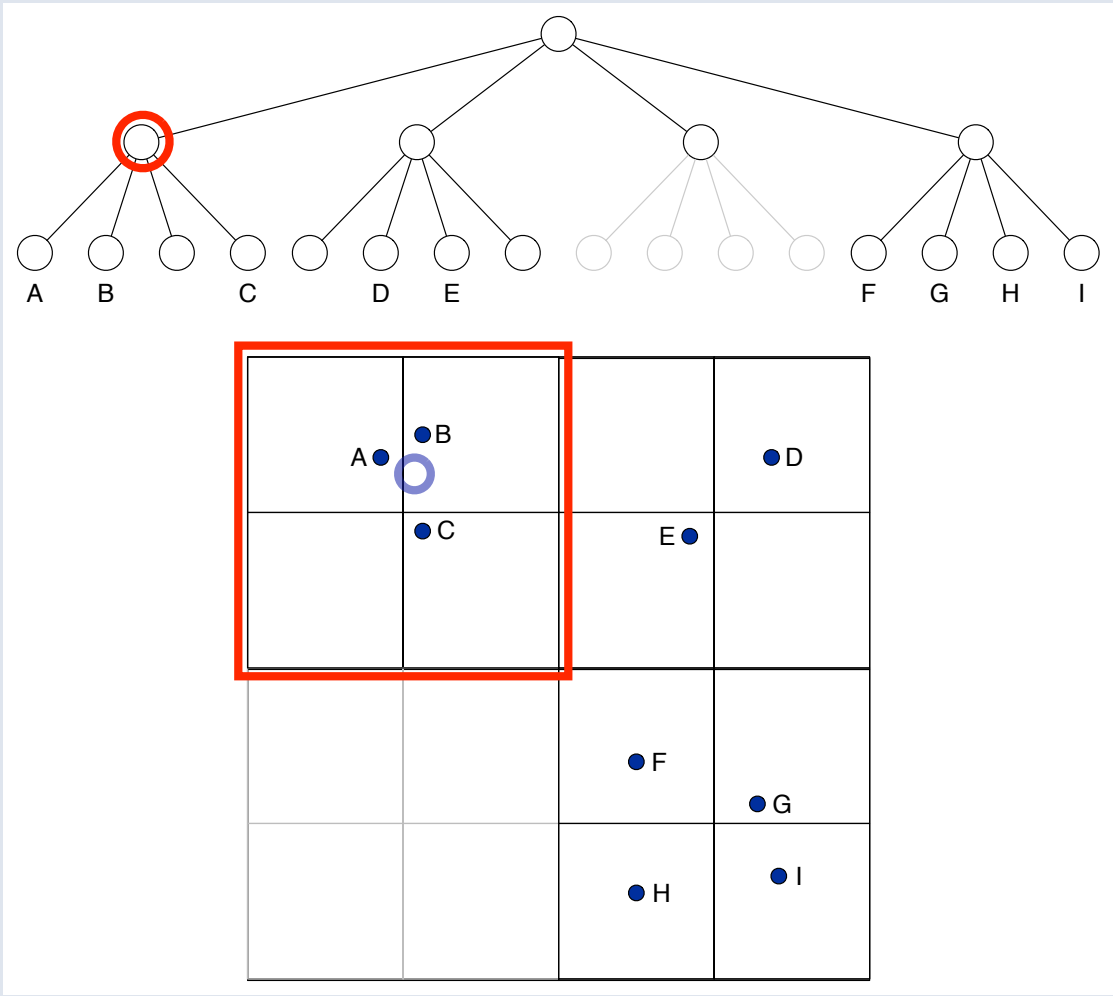
- Approximate such similar interactions by a single interaction:



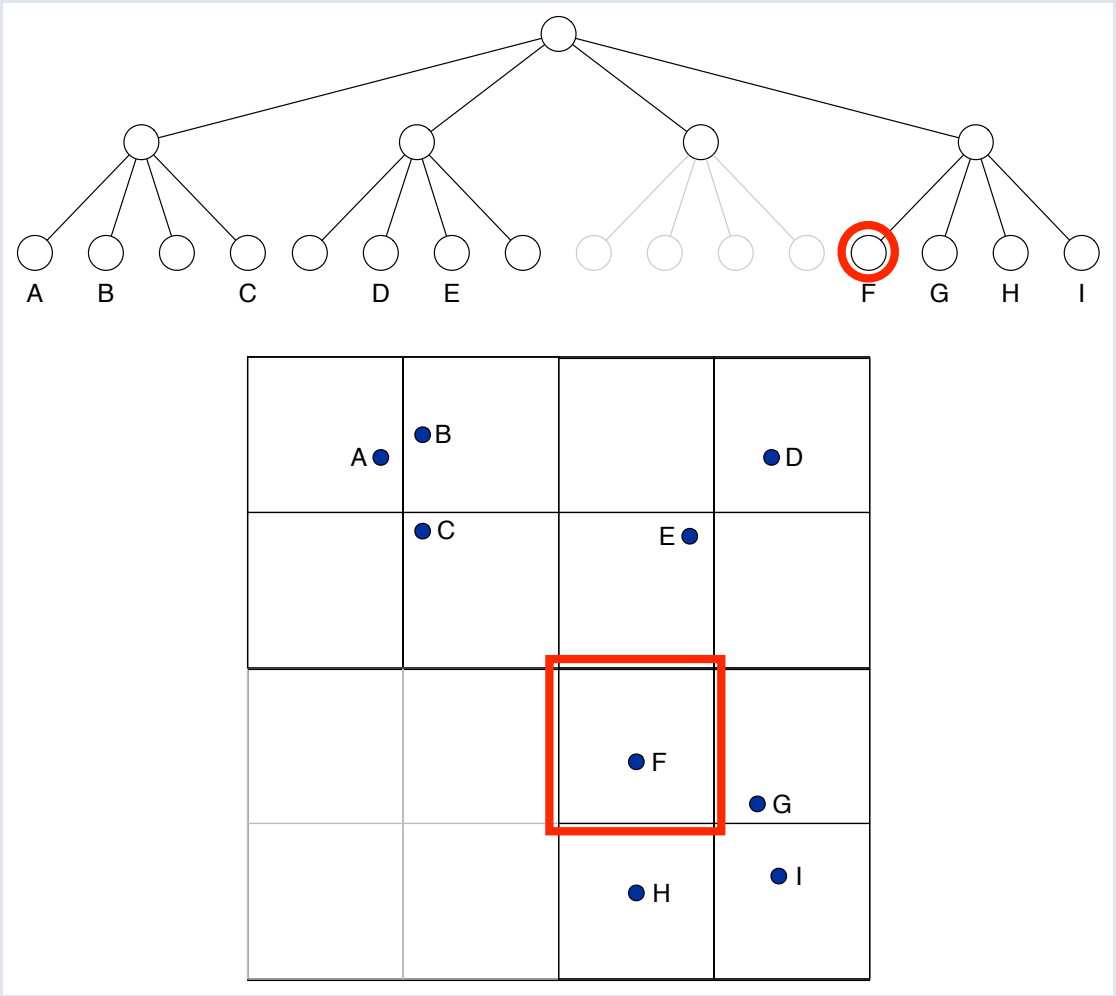
Quadtree



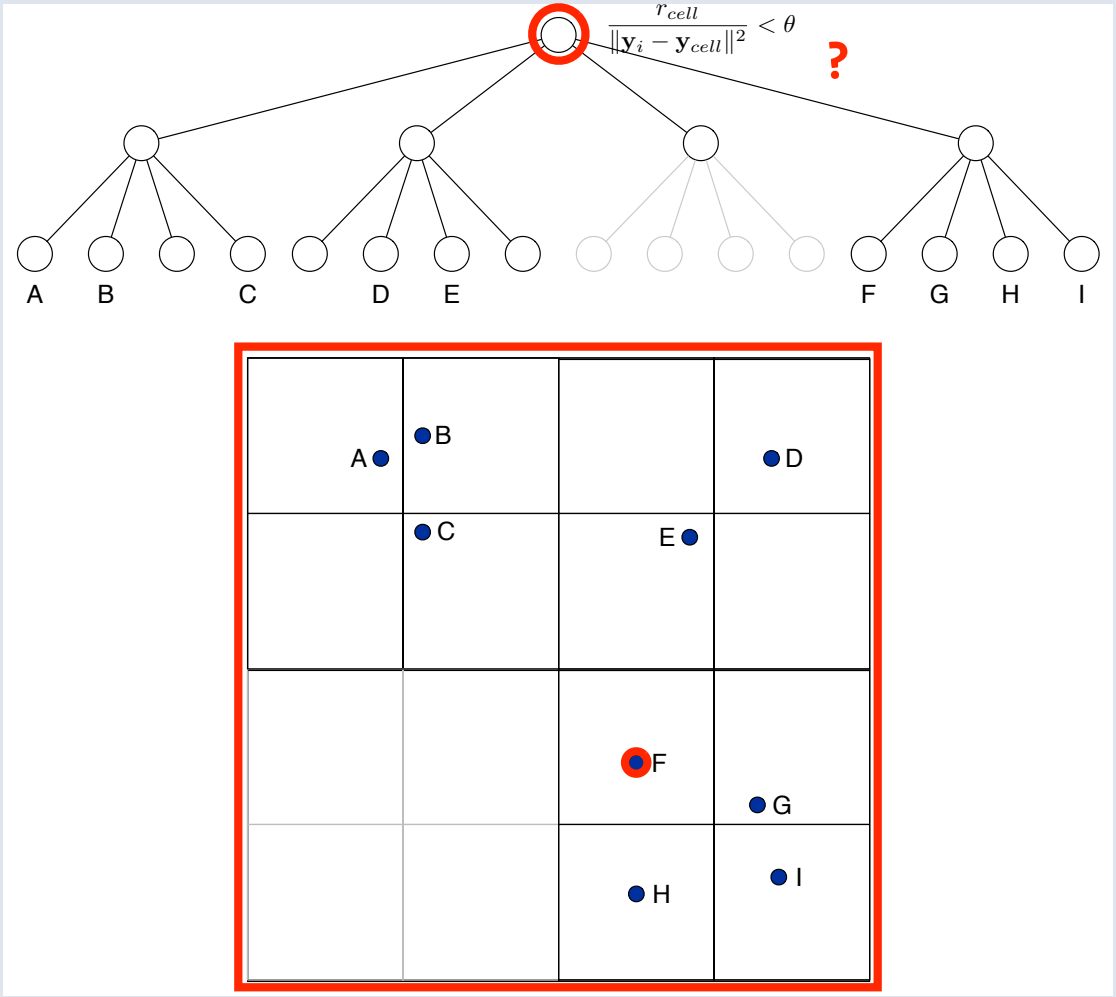
Quadtree



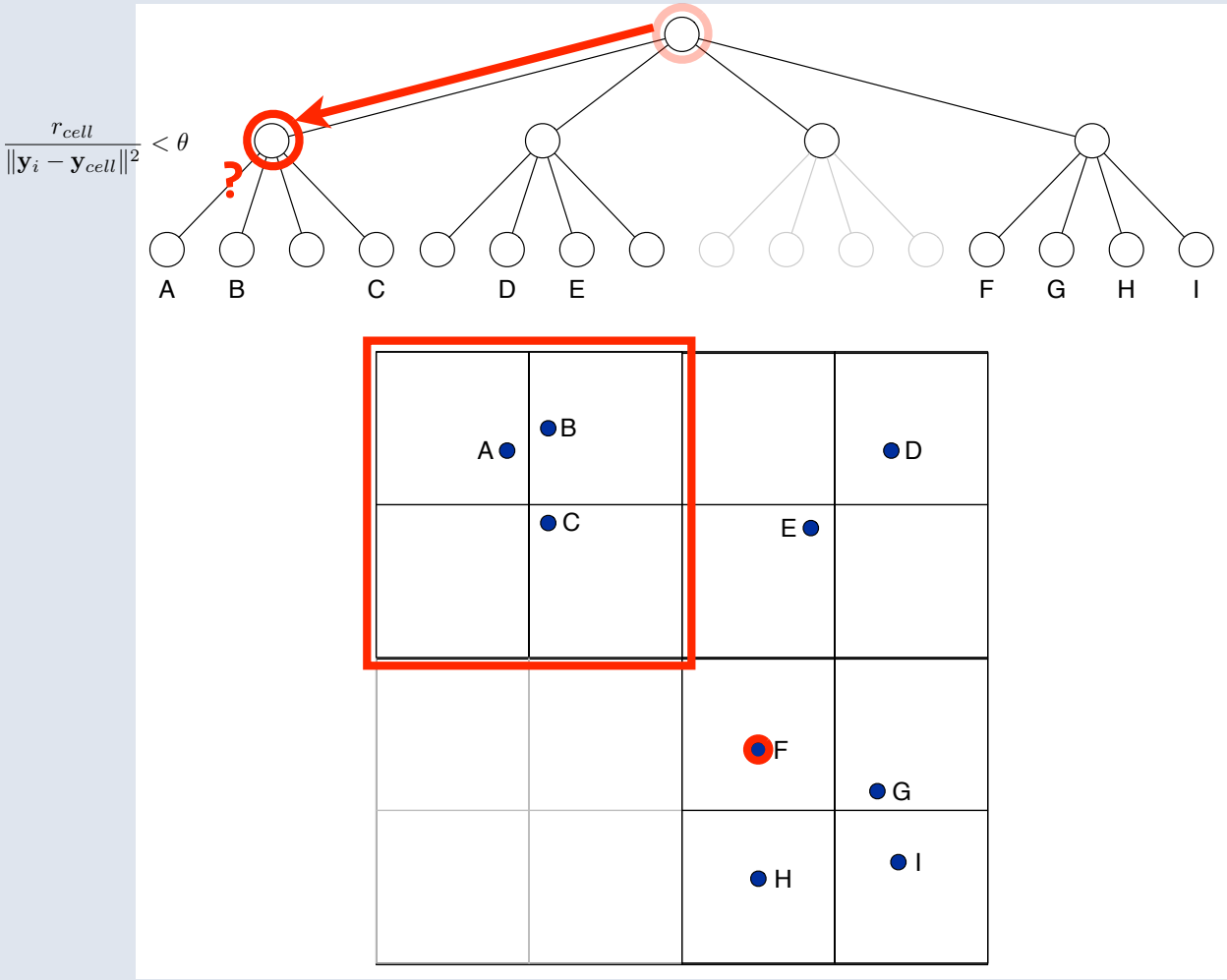
Quadtree



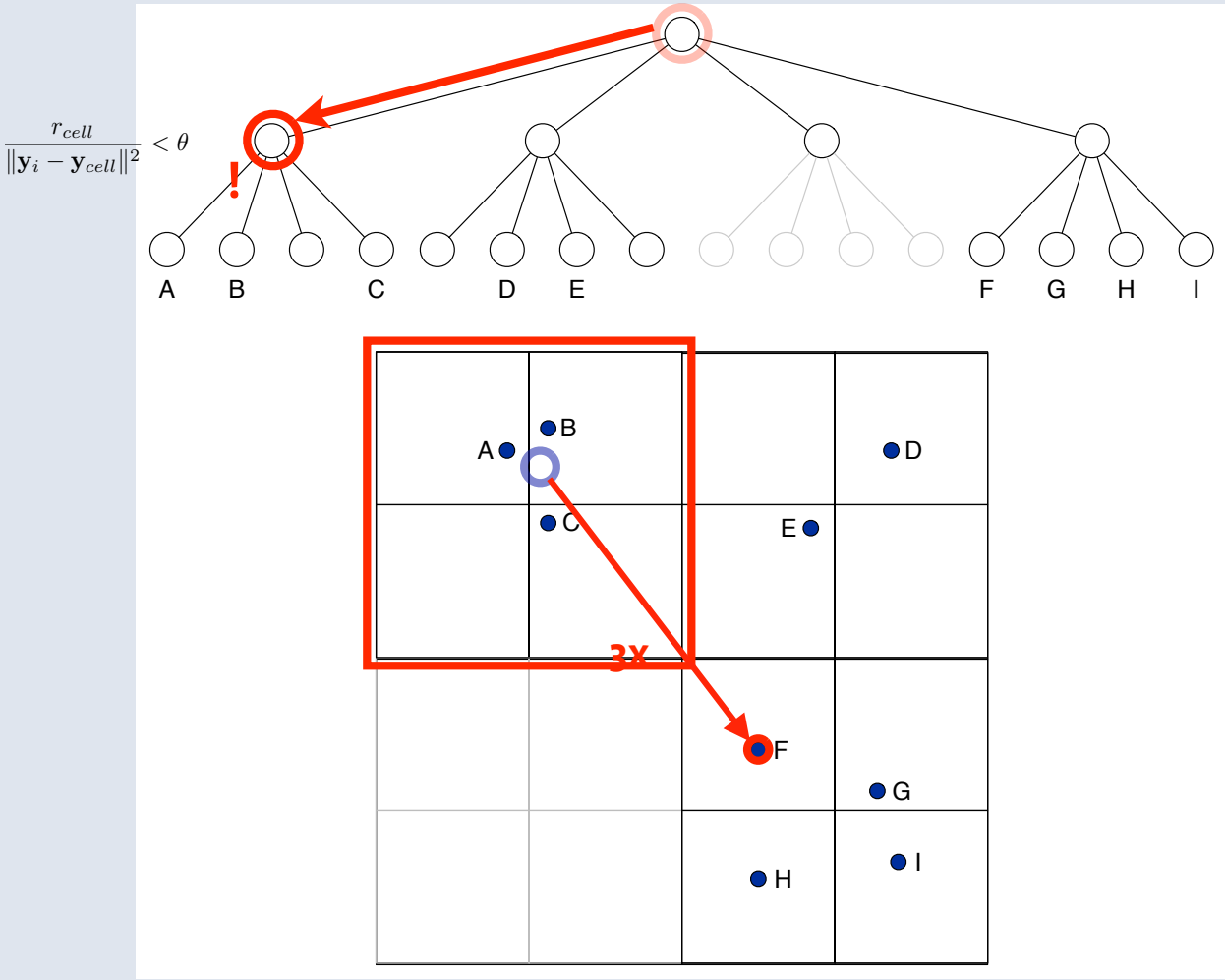
Barnes-Hut-SNE



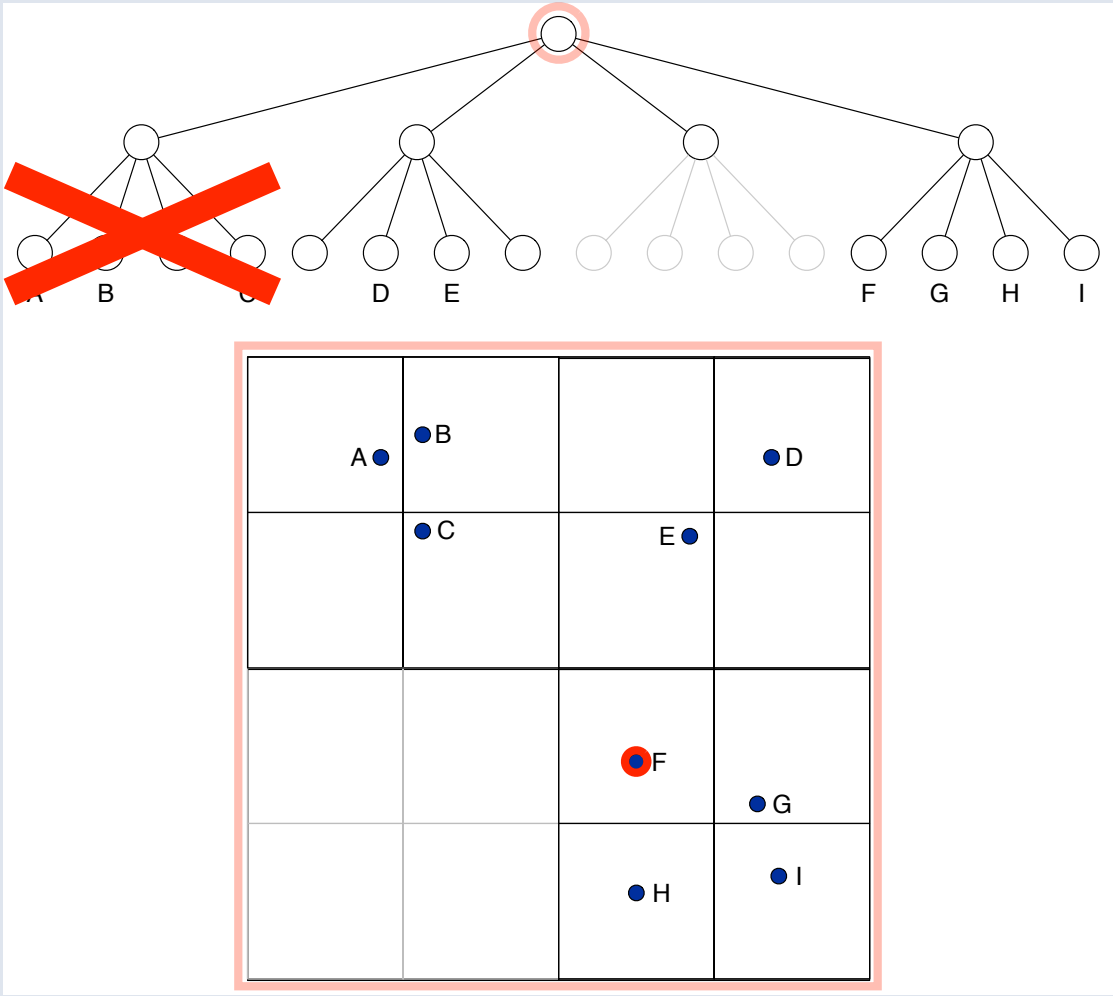
Barnes-Hut-SNE



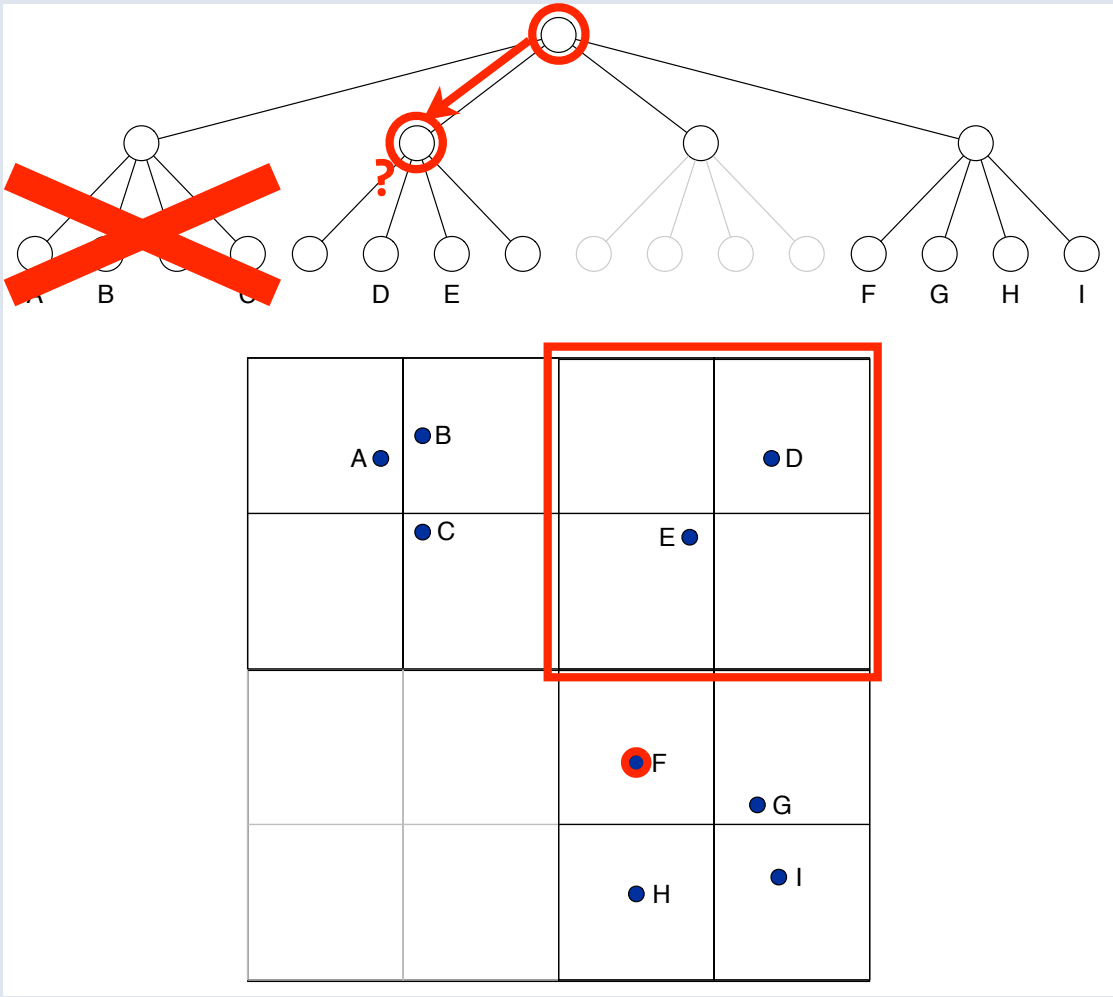
Barnes-Hut-SNE



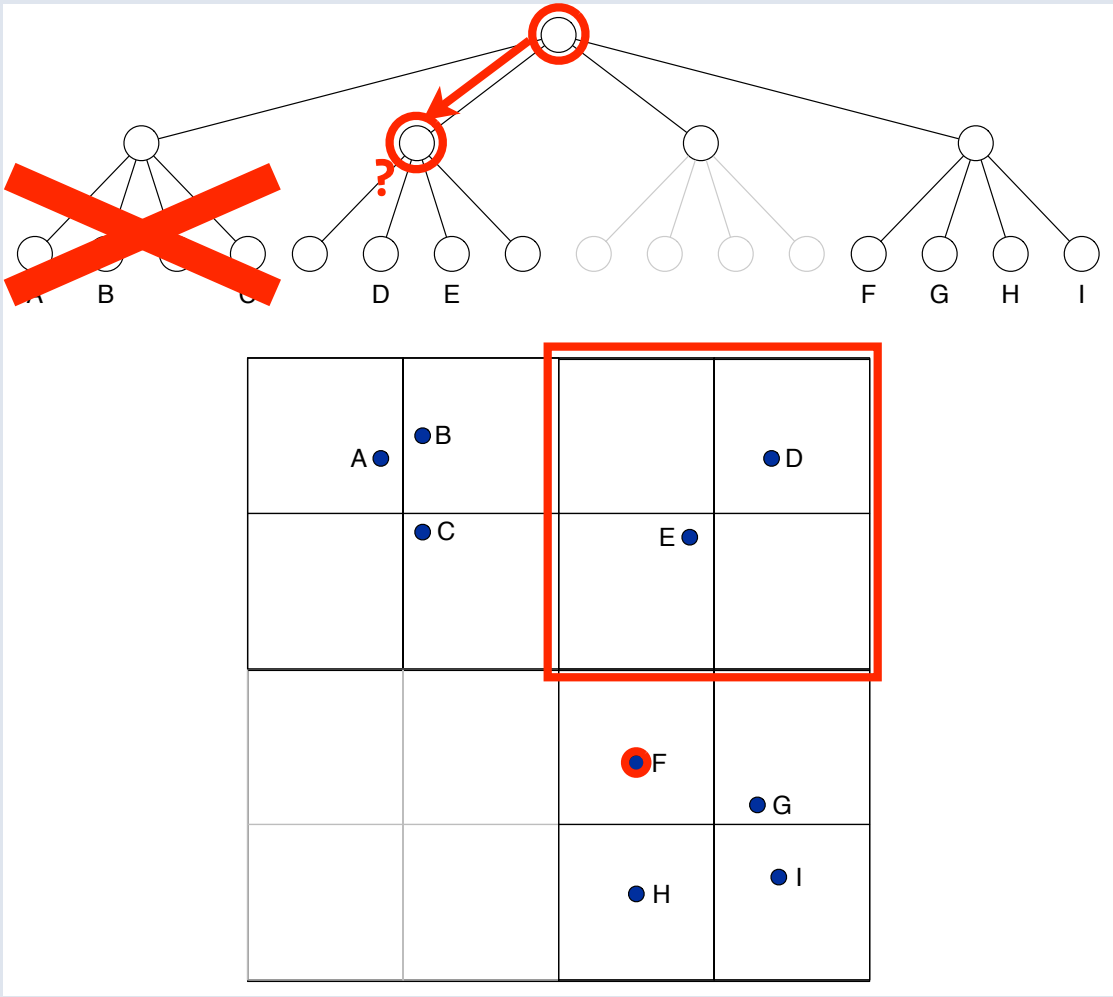
Barnes-Hut-SNE



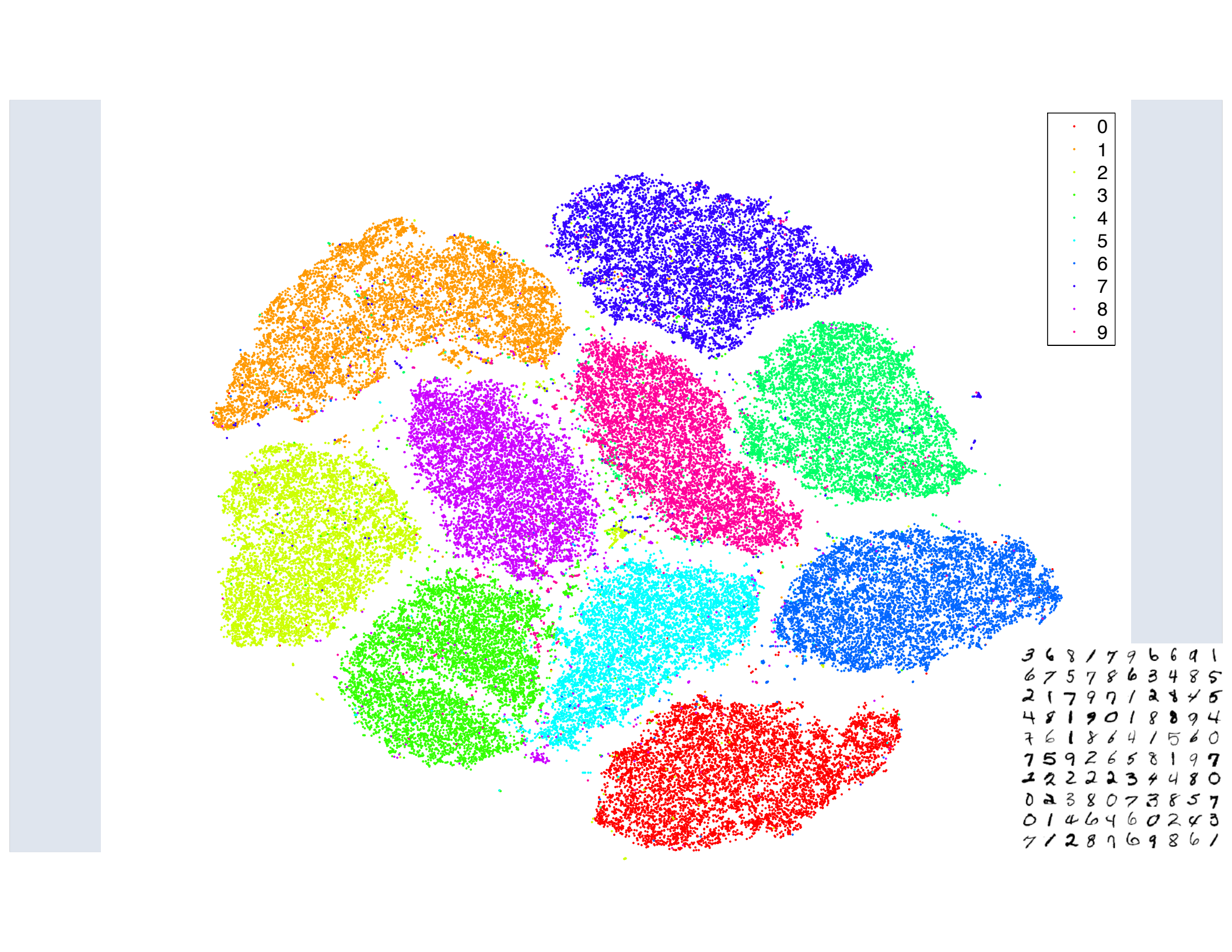
Barnes-Hut-SNE

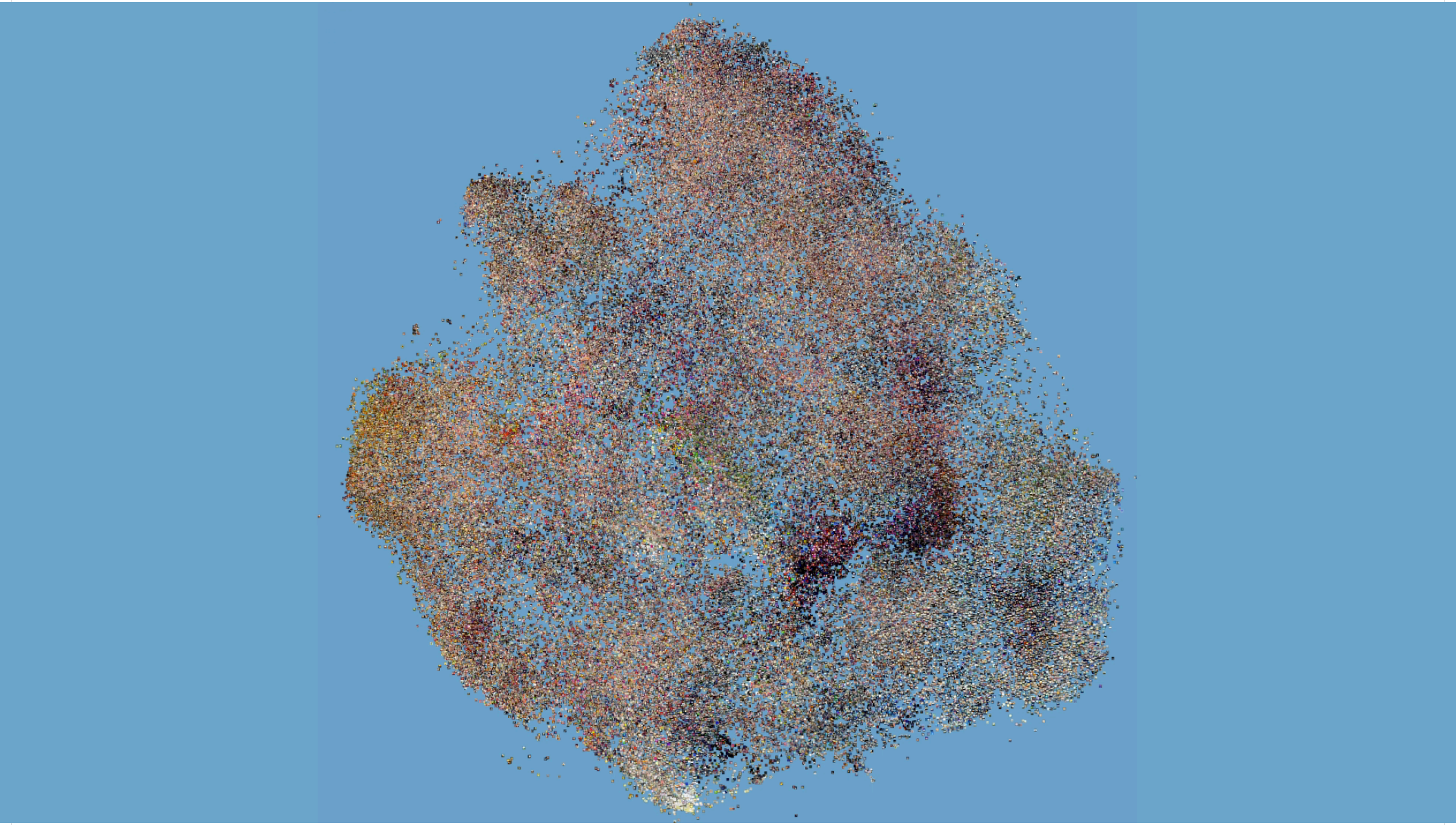


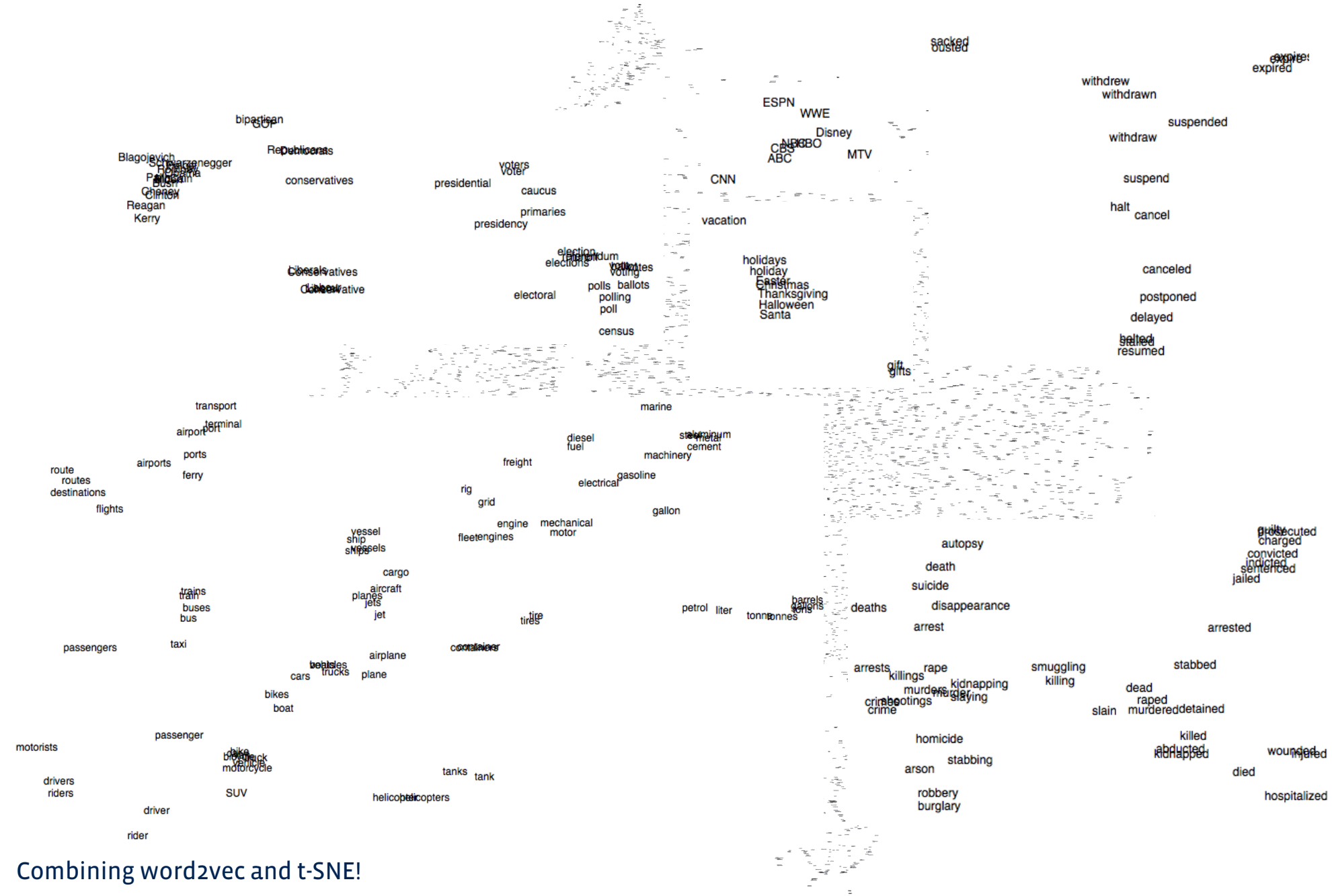
Barnes-Hut-SNE



etcetera...


0
1
2
3
4
5
6
7
8
9





Combining word2vec and t-SNE!

An Important Lesson

- Not all powerful learning algorithms are deep…

- When you do work in this field: always run good baselines!

References

- Reading material:
  - O. Levy and Y. Goldberg. **Neural Word Embedding as Implicit Matrix Factorization**. Advances in Neural Information Processing 27:2177-2185, 2014.
  - A. Joulin, E. Grave, P. Bojanowski, T. Mikolov. **Bag of Tricks for Efficient Text Classification**. arXiv 1607.01759, 2016.
  - L.J.P. van der Maaten and G.E. Hinton. **Visualizing High-Dimensional Data Using t-SNE**. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.
  - A. Jabri, A. Joulin, and L.J.P. van der Maaten. Revisiting Visual Question Answering Baselines. European Conference on Computer Vision, pages 727-739, 2016.
- Source code:
  - Word2vec: https://code.google.com/archive/p/word2vec/
  - FastText: https://github.com/facebookresearch/fastText
  - t-SNE: https://lvdmaaten.github.io/tsne/

facebook

Questions?